IntechOpen

Numerical Simulations

Applications, Examples and Theory

Edited by Lutz Angermann

NUMERICAL SIMULATIONS - APPLICATIONS, EXAMPLES AND THEORY

Edited by **Prof. Lutz Angermann**

Numerical Simulations - Applications, Examples and Theory
http://dx.doi.org/10.5772/901
Edited by Lutz Angermann

Contributors

Magdi Shoucri, Sergey Sysoev, Asdin Aoufi, Gilles Damamme, Alessandro Soprano, Francesco Caputo, Inés Peñuelas, Byoung S. Ham, Teodor Costinel Popescu, Daniela Vasiliu, Nicolae Vasiliu, Constantin Calinoiu, Oleg Yurtcev, Yuri Bobkov, Bin Lin, Feng Liu, Xiaofeng Zhang, Liping Liu, Xueming Zhu, J. Nathan Kutz, Y.B. Guo, Giovanni Bruna, Lyudmila Ryabicheva, Dmytro Usatyuk, Lutz Angermann, Vasyl Vasylyovych Yatsyk, Sergiu Gabriel Spinu, Gheorghe Frunza, Emanuel Diaconescu, Jean-Luc Autran, Daniela Munteanu, Oksana Shpigunova, Anatoly Glazunov, Anne Humeau, Edite Figueiras, Luis F. Requicha Ferreira, Frits F.M. De Mul, Igor Uimanov, Viktorija Grigaitiene, Romualdas Kezelis, Vitas Valincius, Justin Steven Prentice

Notice

First published in Croatia, 2011 by INTECH d.o.o.
eBook (PDF) Published by IN TECH d.o.o.
Place and year of publication of eBook (PDF): Rijeka, 2019.
IntechOpen is the global imprint of IN TECH d.o.o.
Printed in Croatia

Legal deposit, Croatia: National and University Library in Zagreb

Additional hard and PDF copies can be obtained from orders@intechopen.com

Numerical Simulations - Applications, Examples and Theory
Edited by Lutz Angermann

p. cm.

ISBN 978-953-307-440-5

eBook (PDF) ISBN 978-953-51-5555-3

Meet the editor

Lutz Angermann is Professor of Numerical Mathematics in the Mathematical Institute of the University of Technology at Clausthal (Germany) since 2001. His research is concerned with the mathematical analysis of numerical algorithms for partial differential equations with special interests in finite volume and finite element methods. After the study of Mathematics at the University of Kharkov (Ukraine) he earned a Ph.D. from the University of Technology at Dresden in 1987. The University of Erlangen-Nürnberg awarded him a higher doctoral degree (habilitation) in 1995. From 1998 to 2001, he held the post of an Associate Professor of Numerical Mathematics at the University of Magdeburg. He is the author of about 65 scientific papers, among them two co-authored books on numerical methods for partial differential equations.

Contents

Preface

In the recent decades, numerical simulation has become a very important and successful approach for solving complex problems in almost all areas of human life. This book presents a collection of recent contributions of researchers working in the area of numerical simulations. It is aimed to provide new ideas, original results and practical experiences regarding this highly actual field. The subject is mainly driven by the collaboration of scientists working in different disciplines. This interaction can be seen both in the presented topics (for example, problems in fluid dynamics or electromagnetics) as well as in the particular levels of application (for example, numerical calculations, modeling or theoretical investigations).

The papers are organized in thematic sections on computational fluid dynamics (flow models, complex geometries and turbulence, transport of sediments and contaminants, reacting flows and combustion). Since cfd-related topics form a considerable part of the submitted papers, the present first volume is devoted to this area. The second volume is thematically more diverse, it covers the areas of the remaining accepted works ranging from particle physics and optics, electromagnetics, materials science, electrohydraulic systems, and numerical methods up to safety simulation.

In the course of the publishing process it unfortunately came to a difficulty in which consequence the publishing house was forced to win a new editor. Since the undersigned editor entered at a later time into the publishing process, he had only a restricted influence onto the developing process of book. Nevertheless the editor hopes that this book will interest researchers, scientists, engineers and graduate students in many disciplines, who make use of mathematical modeling and computer simulation. Although it represents only a small sample of the research activity on numerical simulations, the book will certainly serve as a valuable tool for researchers interested in getting involved in this multidisciplinary field. It will be useful to encourage further experimental and theoretical researches in the above mentioned areas of numerical simulation.

Lutz Angermann

Institut für Mathematik, Technische Universität Clausthal,
Erzstraße 1, D-38678 Clausthal-Zellerfeld
Germany

Part 1

Particle Physics and Optics

1

Numerical Simulation of the Bump-on-Tail Instability

Magdi Shoucri
Institut de recherche Hydro-Québec (IREQ), Varennes, Québec J3X1S1, Canada

1. Introduction

Wave-particle interaction is among the most important and extensively studied problems in plasma physics. Langmuir waves and their Landau damping or growth are fundamental examples of wave-particle interaction. The bump-on-tail instability is an example of wave growth and is one of the most fundamental and basic instabilities in plasma physics. When the bump in the tail of the distribution function presents a positive slope, a wave perturbation whose phase velocity lies along the positive slope of the distribution function becomes unstable. The bump-on-tail instability has been generally studied analytically and numerically under various approximations, either assuming a cold beam, or the presence of a single wave, or assuming conditions where the beam density is weak so that the unstable wave representing the collective oscillations of the bulk particles exhibits a small growth and can be considered as essentially of slowly varying amplitude in an envelope approximation (see for instance Umeda *et al*., 2003, Doveil *et al*. 2001, and references therein). Some early numerical simulations have studied the growth, saturation and stabilization mechanism for the beam-plasma instability (Dawson and Shanny, 1968, Denavit and Kruer, 1971, Joyce *et al*., 1971, Nührenberg, 1971). Using Eulerian codes for the solution of the Vlasov-Poisson system (Cheng and Knorr, 1976, Gagné and Shoucri, 1977), it has been possible to present a better picture of the nonlinear evolution of the bump-on-tail instability (Shoucri, 1979), where it has been shown that for a single wave perturbation the initial bump in the tail of the distribution is distorted during the instability, and evolves to an asymptotic state having another bump in the tail of the spatially averaged distribution function, with a minimum of zero slope at the phase velocity of the initially unstable wave (in this way the large amplitude wave can oscillate at constant amplitude without growth or damping). The phase-space in this case shows in the asymptotic state a Bernstein-Greene-Kruskal (BGK) vortex structure traveling at the phase-velocity of the wave (Bernstein *et al*., 1957, Bertrand *et al*., 1988, Buchanan and Dorning, 1995). These results are also confirmed in several simulations (see for instance Nakamura and Yabe, 1999, Crouseilles *et al*., 2009). Since the early work of Berk and Roberts, 1967, the existence of steady-state phase-space holes in plasmas has been discussed in several publications. A discussion on the formation and dynamics of coherent structures involving phase-space holes in plasmas has been presented for instance in the recent works of Schamel, 2000, Eliasson and Shukla, 2006.
There are of course situations where a single wave theory and a weak beam density do not apply. In the present Chapter, we present a study for the long-time evolution of the Vlasov-

Poisson system for the problem of the bump-on-tail instability, for the case when the beam density is about 10% of the total density, which provides a more vigorous beam-plasma interaction and important wave-particle and trapped particles effects. In this case the instability and trapping oscillations have important feedback effects on the oscillation of the bulk. Since the bump in the tail is usually located in the low density region of the distribution function, the Eulerian codes, because of their low noise level, allow an accurate study of the evolution of the bump, and on the transient dynamics for the formation and representation of the traveling BGK structures (for details on the numerical codes see the recent articles in Pohn *et al*., 2005, Shoucri, 2008, 2009). A warm beam is considered, and the system length L is greater than the wavelength of the unstable mode λ. In this case growing sidebands develop with energy flowing to the longest wavelengths (inverse cascade). This inverse cascade is characteristic of 2D systems (Knorr, 1977). Oscillations at frequencies below the plasma frequency are associated with the longest wavelengths, and result in phase velocities above the initial beam velocity, trapping and accelerating particles to higher velocities. The electric energy of the system is reaching in the asymptotic state a steady state with constant amplitude modulated by the persistent oscillation of the trapped particles, and of particles which are trapped, untrapped and retrapped. A similar problem has been recently studied in Shoucri, 2010. In the present chapter, we shall consider a larger simulation box, capable of resolving a broader spectrum. Two cases will be studied. A case where a single unstable mode is initially excited, and a case where two unstable modes are initially excited. Differences in the results between these two cases will be pointed out.

The transient dynamics of the Vlasov-Poisson system is sensitive to grid size effects (see, for instance, Shoucri, 2010, and references therein). Numerical grid size effects and small time-steps can have important consequences on the number and distribution of the trapped particles, on kinetic microscopic processes such as the chaotic trajectories which appear in the resonance region at the separatrix of the vortex structures where particles can make periodic transitions from trapped to untrapped motion. Usually during the evolution of the system, once the microstructure in the phase-space is reaching the mesh size, it is smoothed away by numerical diffusion, and is therefore lost. Larger scales appear to be unaffected by the small scale diffusivity and appear to be treated with good accuracy. This however has consequences on smoothing out information on trapped particles, and modifying some of the oscillations associated with these trapped particles, and with particles at the separatrix region of the vortex structures which evolve periodically between trapping and untrapping states. These trapped particles play an important role in the macroscopic nonlinear oscillation and modulation of the asymptotic state, and require a fine resolution phase-space grid and a very low noise code to be studied as accurately as possible (Califano and Lantano, 1999, Califano *et al*., 2000, Doveil *et al*., 2001, Valentini *et al*., 2005, Shoucri, 2010).

The transient dynamics of the Vlasov-Poisson system is also sensitive to the initial perturbation of the system. Two cases will be considered in this chapter in the context of the bump-on-tail instability. A case where a single unstable mode is initially excited, and a case where two unstable modes are initially excited. In the first case, the system reaches in a first stage a BGK traveling wave, which in this case with $L > \lambda$ is only an intermediate state. Growing sidebands develop which disrupt the BGK structure and the system evolves in the end to a phase-space hole which translates as a cavity-like structure in the density plot. In the case where two initially unstable modes are excited, the electric energy decays rapidly after the initial growth and the vortices formed initially are unstable, and the phase-space evolves rapidly to a structure with a hole. In both cases energy is transferred by inverse

cascade to the longest wavelengths available in the system. A more important heating of the tail is observed in this second case.

2. The relevant equations

The relevant equations are the 1D Vlasov equation for the electron and ion distribution functions $f_e(x,\upsilon_e,t)$ and $f_i(x,\upsilon_i,t)$, coupled to the Poisson equation. These equations are written in our normalized units:

$$\frac{\partial f_{e,i}}{\partial t}+\frac{\partial f_{e,i}}{\partial x}\mp\frac{1}{m_{e,i}}E_x\frac{\partial f_{e,i}}{\partial \upsilon_{e,i}}=0 \tag{1}$$

$$\frac{\partial^2\varphi}{\partial x^2}=-(n_i(x)-n_e(x))\;;\; n_{e,i}(x)=\int_{-\infty}^{+\infty} f_{e,i}(x,\upsilon_{e,i})d\upsilon_{e,i} \tag{2}$$

$$E_x=-\frac{\partial\varphi}{\partial x} \tag{3}$$

Time t is normalized to the inverse electron plasma frequency ω_{pe}^{-1}, velocity is normalized to the electron thermal velocity $\upsilon_{the}=\sqrt{T_e/m_e}$ and length is normalized to the Debye length $\lambda_{D_e}=\upsilon_{the}/\omega_{pe}$. In our normalized units, $m_e=1$ and $m_i=M_i/M_e$. Periodic boundary conditions are used. These equations are discretized on a grid in phase-space and are solved with an Eulerian code, by applying a method of fractional step which has been previously presented in the literature (Cheng and Knorr, 1976, Gagné and Shoucri, 1977, Shoucri, 2008, 2009). The distribution function for a homogeneous electron beam-plasma system, with an electron beam drifting with a velocity υ_d relative to a stationary homogeneous plasma is given by:

$$f_e(\upsilon_e)=\frac{n_p}{\sqrt{2\pi}}e^{-\frac{1}{2}\upsilon_e^2}+\frac{n_b}{\sqrt{2\pi}\upsilon_{thb}}e^{-\frac{1}{2}\frac{(\upsilon_e-\upsilon_d)^2}{\upsilon_{thb}^2}} \tag{4}$$

The electron beam thermal spread is $\upsilon_{thb}=0.5$ and the beam velocity is $\upsilon_d=4.5$. The ion distribution function in our normalized units is given by:

$$f_i(\upsilon_i)=\frac{n_i}{\sqrt{2\pi}\upsilon_{thi}}e^{-\frac{1}{2}\upsilon_i^2/\upsilon_{thi}^2} \tag{5}$$

We take for the electron plasma density $n_p=0.9$ and for the electron beam density $n_b=0.1$ for a total density of 1. This high beam density will cause a strong beam-plasma instability to develop. We take $n_i=1.$, $T_e/T_i=1$, $m_e/m_i=1./1836$. In our normalized units $\upsilon_{thi}=\sqrt{T_i m_e/T_e m_i}$. We use a time-step $\Delta t=0.002$. The length of the system in the present simulations is $L=80\text{x}2\pi/3=167.552$.

3. Excitation of the mode n=8 with k=0.3

We perturb the system initially with a perturbation such that:

$$f_e(x,\upsilon_e) = f(\upsilon_e)(1+\varepsilon\cos(kx)) \tag{6}$$

with $\varepsilon = 0.04$ and with $k = n\frac{2\pi}{L}$, and $f_e(\upsilon_e)$ is given in Eq.(4). We consider the case where $k = 0.3$, and the approximate initial frequency response of the system will be $\omega^2 \approx 1 + 3k^2$, or $\omega \approx 1.127$ (nonlinear solutions can give slightly different results) , with a phase velocity of the wave $\omega / k \approx 3.756$. This phase velocity corresponds to a velocity where the initial distribution function in Eq.(4) has a positive slope. Hence the density perturbation in Eq.(6) will lead to an instability. The mode $k = 0.3$ corresponds to the mode with *n*=8, in which case unstable sidebands can grow (the length of the system is bigger with respect to the wavelength of the excited oscillation). We use a space-velocity grid of 1024x2400 for the electrons, with extrema in the electron velocity equal to ± 8. The recurrence time in this case is $\tau = \frac{2\pi}{k\Delta\upsilon} \approx 3140$. We use a space-velocity grid of 1024x800 for the ions.

Unstable sidebands are growing from round-off errors. Fig.(1) presents the time evolution of the electric energy, showing growth, saturation and trapped particle oscillations until around a time *t*=700. Figs.(2a,b) show the contour plot and a three-dimensional view at *t*=680 of the distribution function showing the formation of a stable structure of eight vortices, corresponding to the initially unstable *n*=8 mode. The frequency spectrum of the mode *n*=8 at this stage of the evolution of the system shows a dominant frequency at $\omega = 1.0258$ (see Fig.(19b)), corresponding to a phase velocity $\upsilon \approx 3.42$. This corresponds to the velocity at which the center of the BGK structure of in Fig.(2a) is traveling. The spatially averaged distribution function $F_e(\upsilon_e)$ in Fig.(3) is calculated from:

$$F_e(\upsilon_e) = \frac{1}{L}\int_0^L f(x,\upsilon_e)dx \tag{7}$$

The spatially averaged distribution function at this stage of the evolution has evolved from the initial bump-on-tail configuration (full curve in Fig.(3)), to a shape having another bump-on-tail configuration, with a minimum at $\upsilon \approx 3.42$, which corresponds to the phase velocity of the dominant *n*=8 mode at this stage. So the *n*=8 mode is reaching at this early stage a constant amplitude modulated by the oscillation of the trapped particles (see Fig.(19a)), with its phase velocity at the local minimum of the spatially averaged distribution function. During this phase of the evolution the spectrum of the *n*=8 mode in Fig.(19b) shows also the presence of a frequency at $\omega = 1.3134$, which corresponds to a phase velocity $\omega / k = 4.378$, at the local maximum appearing around $\upsilon \approx 4.4$ in the spatially averaged distribution function in Fig.(3). Above $\upsilon \approx 4.4$, the spatially averaged distribution function in Fig.(3) shows a small oscillation with a local minimum due to the trapped population which is apparent above the vortices in Fig.(2a). Fig.(4a) and Fig.(4b) show respectively the electric field and the electron density profiles at *t*=680.

Then for $t > 700$ there is a rapid decrease in the electric energy down to a constant value, (see Fig.(1)). This is caused by the growing sidebands who have reached a level where they

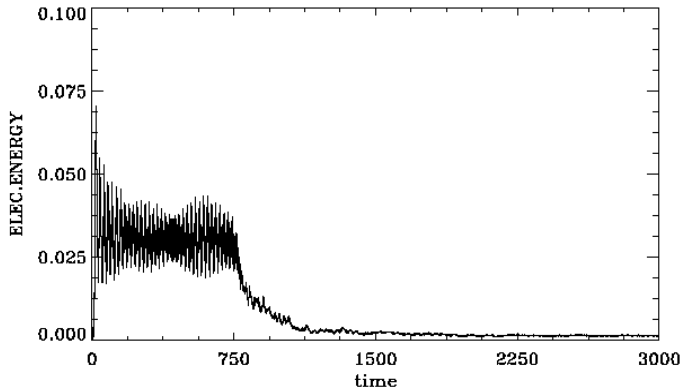
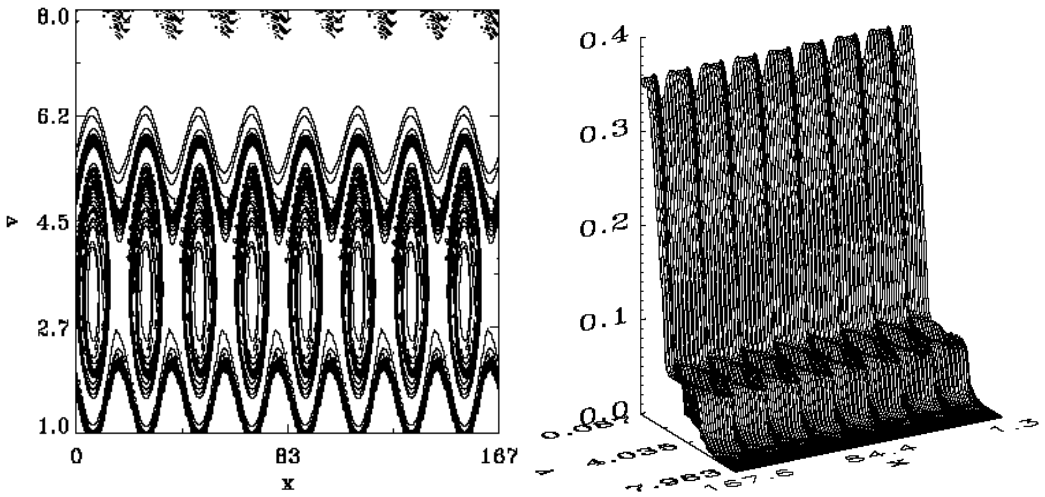


Fig. 1. Time evolution of the electric field energy



(a) (b)

Fig. 2. (a) Contour plot of the distribution function, *t*=680, (b) Three-dimensional view of the distribution function, *t*=680

interact with the eight vortices BGK structure formed. There is a rapid fusion of the vortices into a single vortex, with energy cascading to the longest wavelengths associated with the system, a process characteristic of 2D systems (Knorr,1977). There is a heating of the distribution function, with an elongation of the tail of the distribution. Figs.(5-7) show the sequence of events in the evolution of the phase-space from the eight vortices BGK structure of Fig.(2a) to a single hole structure in Fig.(7o). Fig.(5a) shows at *t*=760 the disruption of the symmetry of the eight vortices structure. Some details are interesting. We note in Fig.(5a) two small vortices, centered around $x \approx 30$ and $x \approx 155$, extending an arm embracing the vortex on their right. We magnify in Fig.(5b) the small vortex centered around $x \approx 30$. Fig.(6a) shows the phase- space at *t*=780. Between Fig.(5a) and Fig.(6a), there is a time delay of 20, in which the structure moves a distance of about $3.42 \text{x} 20 \approx 68$. The small vortex centered at $x \approx 30$ in Figs.(5a,b) has now moved to the position $x \approx 98$ in Fig.(6a).

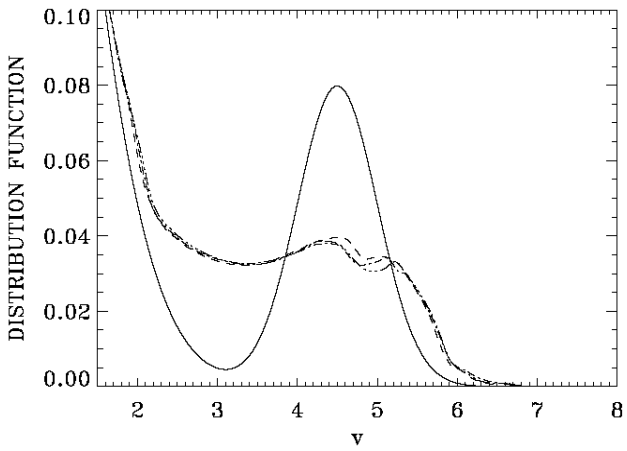


Fig. 3. Spatially averaged distribution function at t=0 (full curve), t=660 (dashed curve), t=680 (dashed-dotted curve), t=700 (dashed-three-dotted curve)

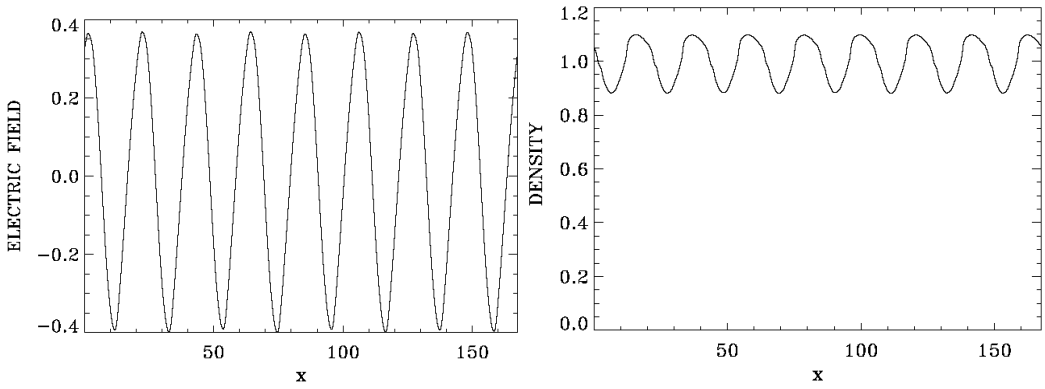


Fig. 4. (a) Electric field profile at t=680, (b) Electron density profile at t=680

We show in more details in Fig.(6b) these vortices structure which now extend their arms to embrace the neighbouring vortices, both to the right and to the left.

We present in Fig.(7a-o) the sequence of evolution of the phase-space, leading to the formation of a single hole structure in Fig.(7o). Note in Fig.(7g) how the tail of the distribution function has shifted to higher velocities. The sequences in Fig.(7h-o) showing the fusion of the final two vortices is interesting. Fig.(7i) shows that one of the two holes is taking a satellite position with respect to the other one, and then is elongated to form an arm around the central vortex. It appears that the satellite vortex is following a spiral structure around the central vortex, possibly following the separatrix. Fig.(8) shows a 3D view of the distribution function at t=2980, corresponding to the results in Fig.(7o). The center of the hole in the phase-space is traveling at a velocity around ≈ 4.8, which is the phase velocity of the dominant modes in Figs.(12-22), as it will be discussed later on. Note the difference in the structure of the electron distribution function between Fig.(8) and Fig.(2b). In Fig.(8), there is a cavity like structure which extends deep in the bulk and which propagates as a solitary like structure in the phase-space at the phase velocity of the hole.

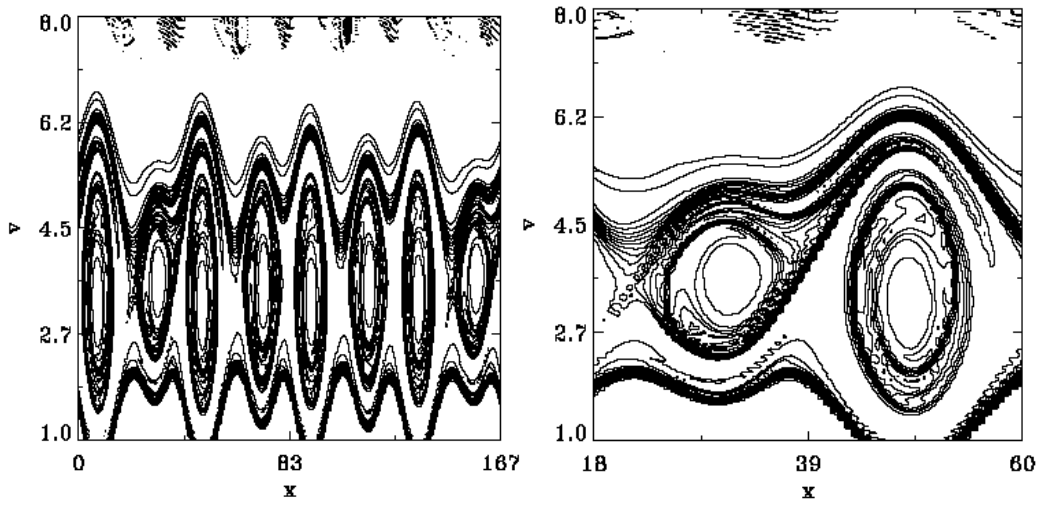


Fig. 5. (a) Contour plot of the distribution function, t=760, (b) same as Fig.5a, figure centered at x=39

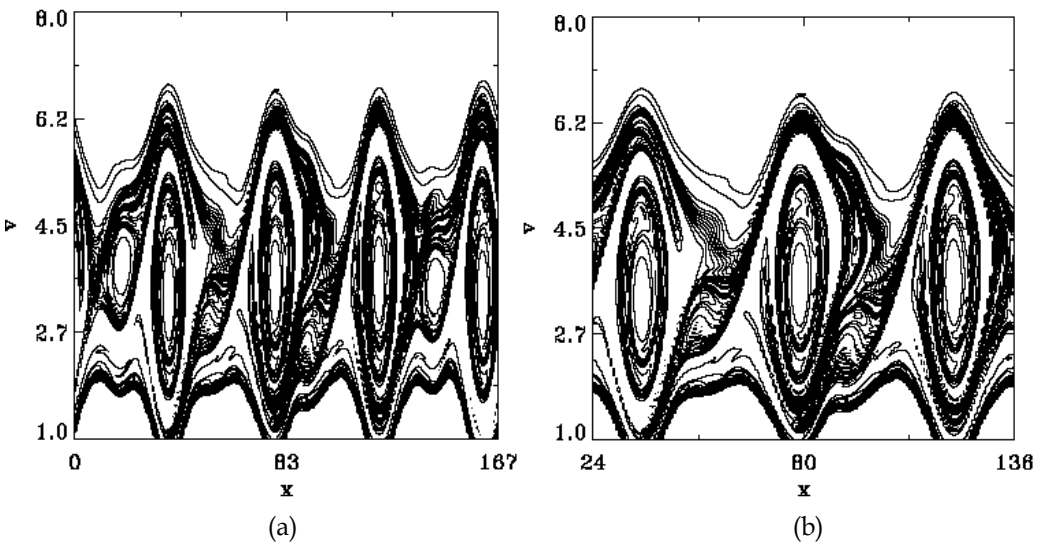


Fig. 6. (a) Contour plot of the distribution function, t=780, (b) same as Fig.6a, centered at x=80

Figs.(9) shows the spatially averaged distribution function for electrons at $t = 2980$, calculated using Eq.(7). Fig.(10) shows the equivalent plot of the ions calculated from an equation equivalent to Eq.(7). Fig.(9) seems to indicate the formation of a plateau. The ion distribution function in Fig.(10) is essentially the same as the initial one at $t = 0$. In Fig.(11a) we plot on a logarithmic scale what appears to be the region of a plateau in Fig.(9). We see in Fig.(11a) the distribution function is decaying slowly, showing an inflexion point around $\upsilon \approx 3.7$, and another one around $\upsilon \approx 4.8$. Fig.(11b) shows on a logarithmic scale a plot of the

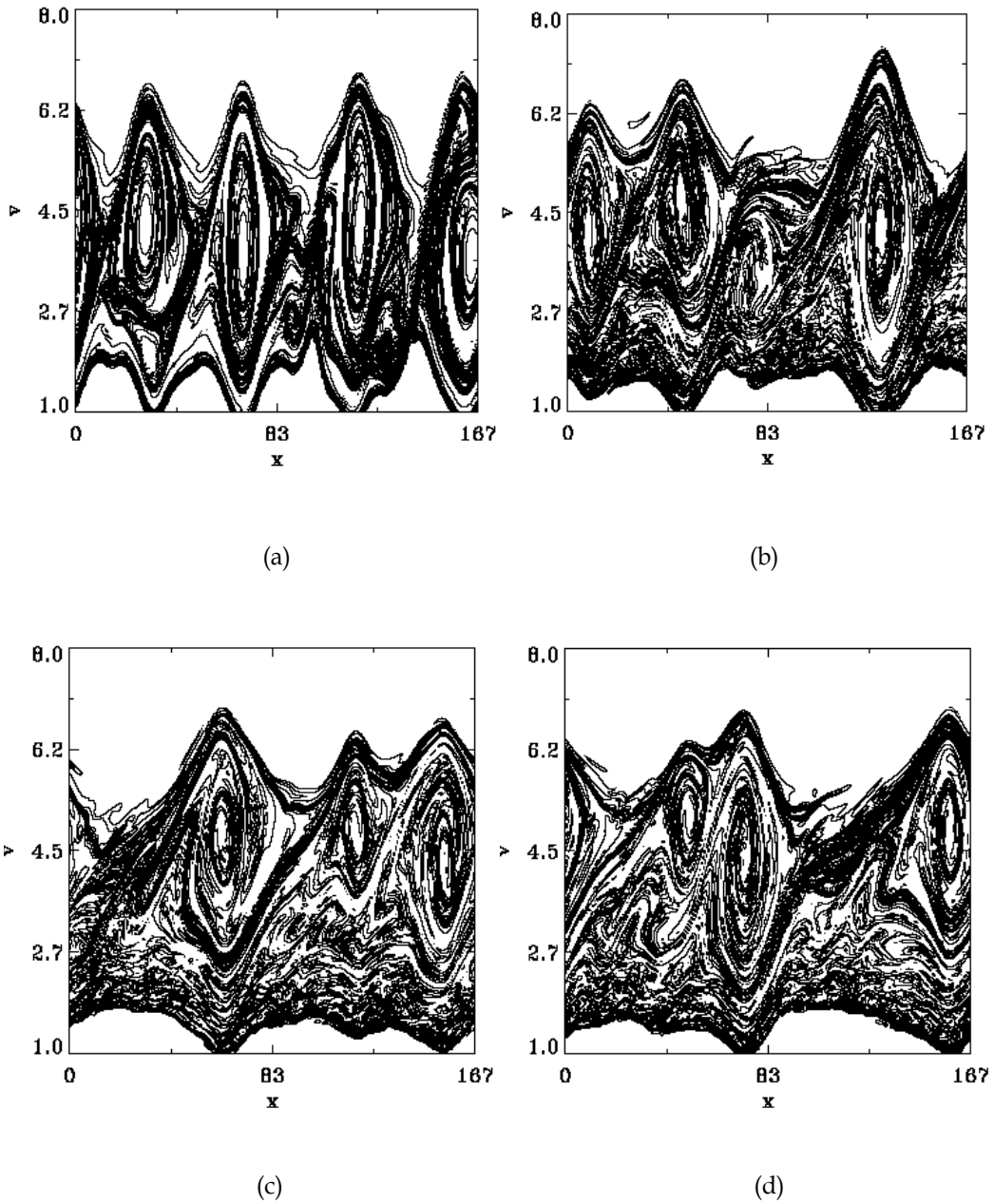


Fig. 7. (a) Contour plot of the distribution function, t=800,
(b) Contour plot of the distribution function, t=1040,
(c) Contour plot of the distribution function, t=1100,
(d) Contour plot of the distribution function, t=1120.

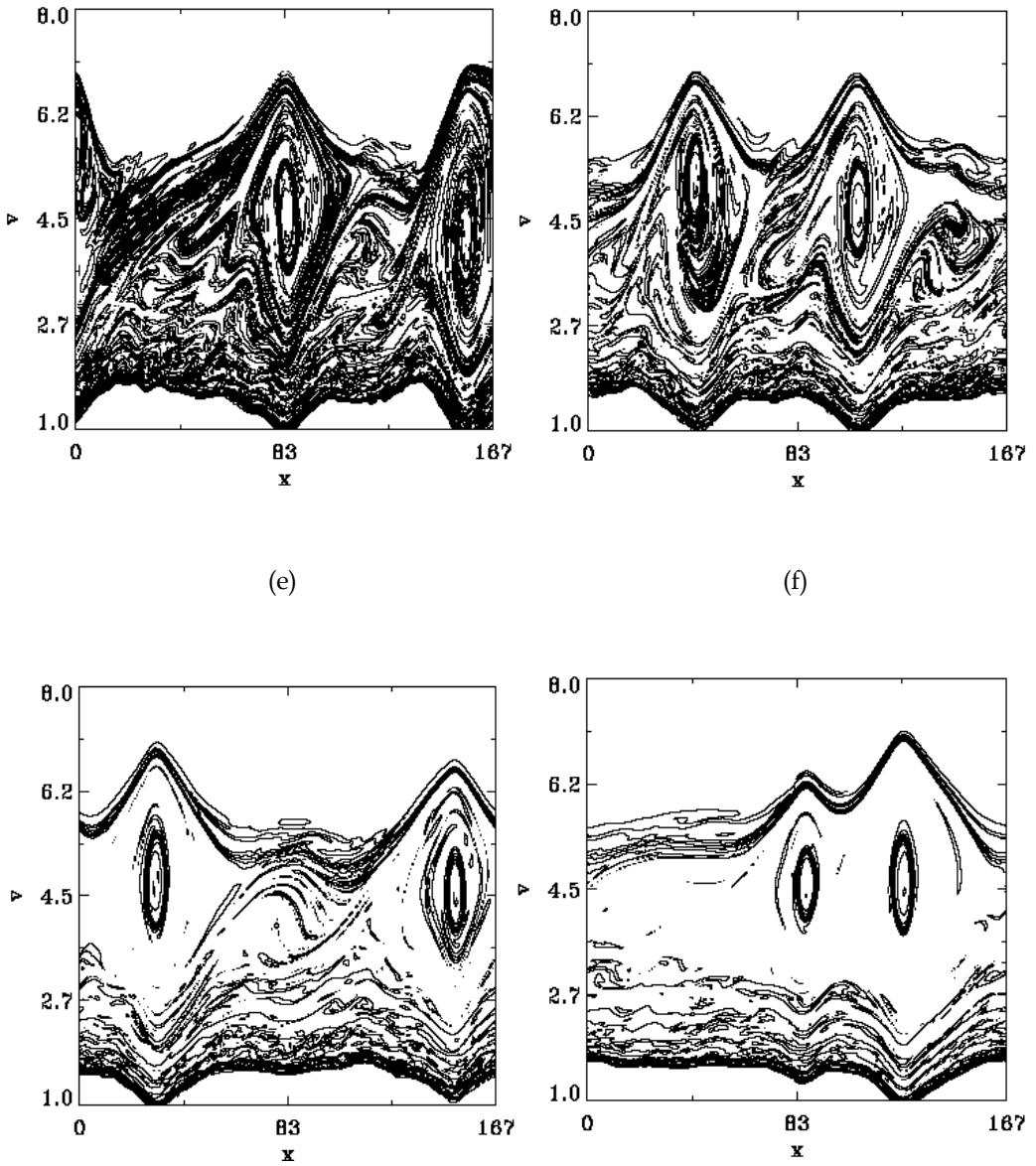


Fig. 7. (e) Contour plot of the distribution function, t=1140,
(f) Contour plot of the distribution function, t=1400,
(g) Contour plot of the distribution function, t=1600,
(h) Contour plot of the distribution function, t=1800.

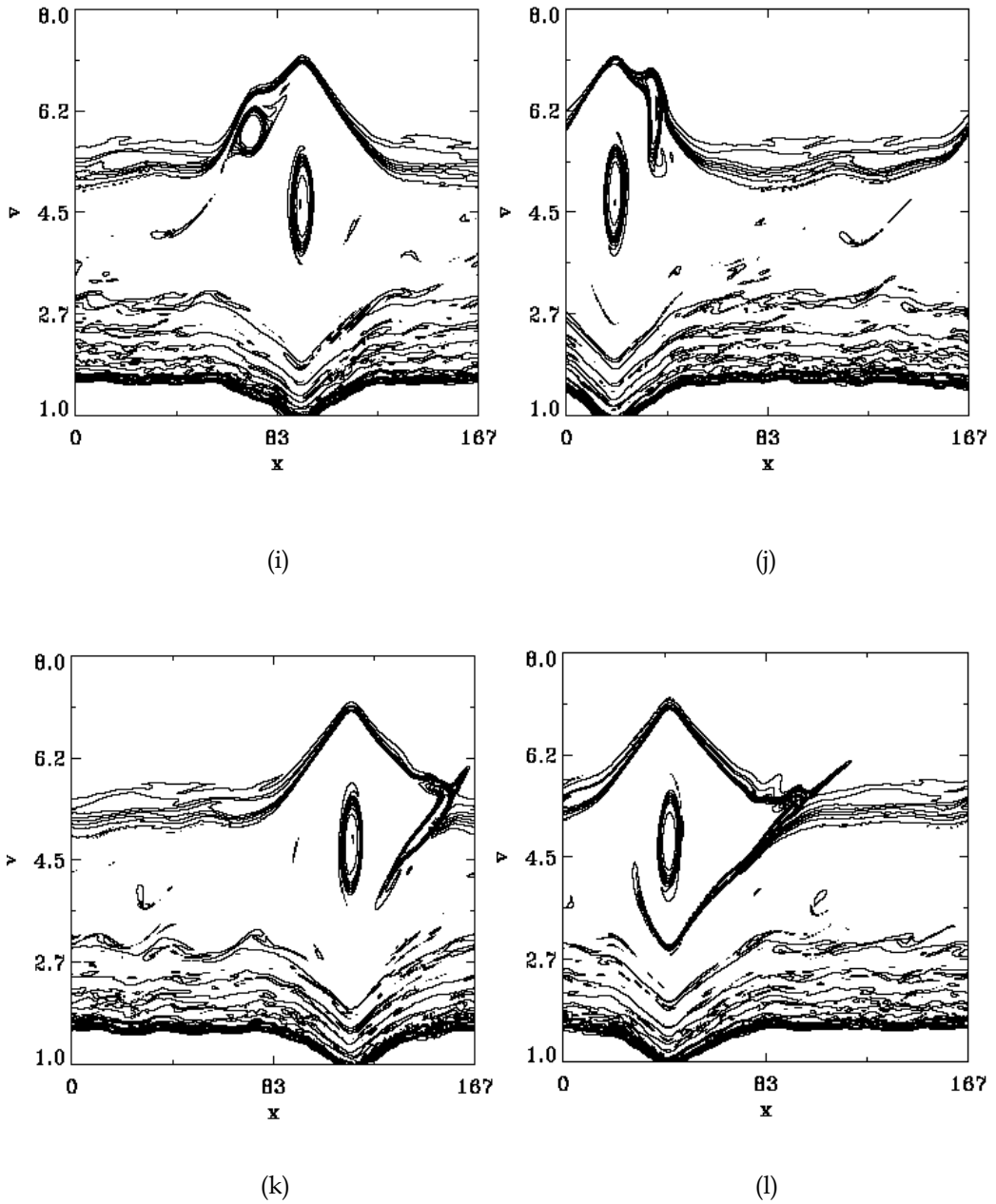
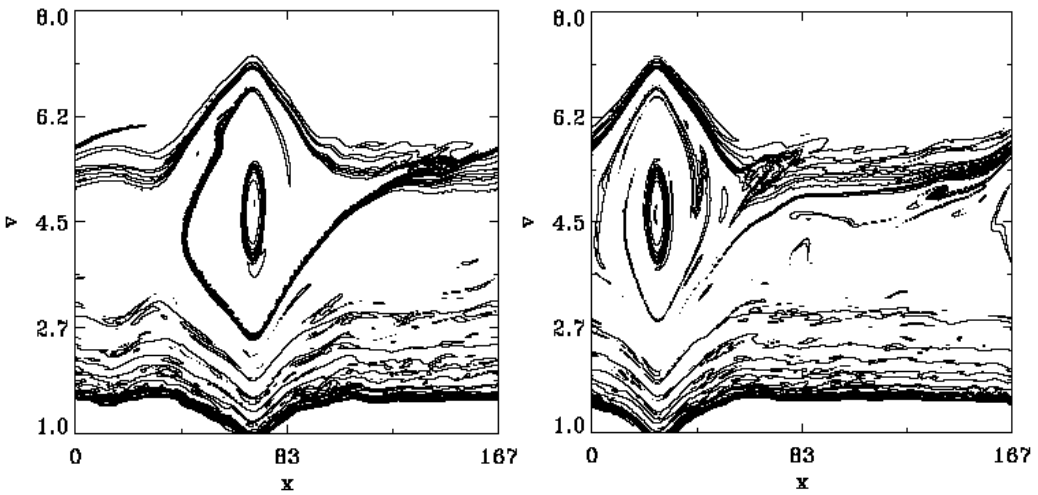
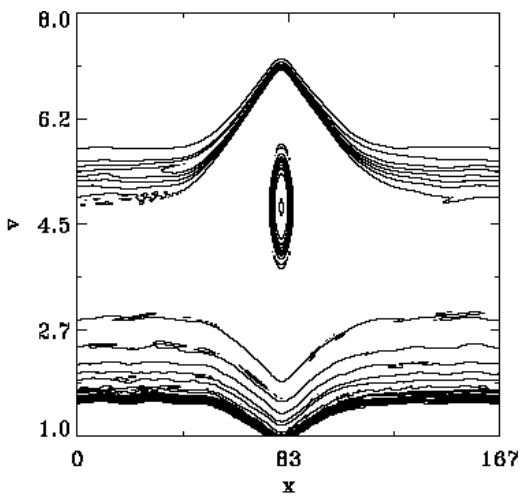


Fig. 7. (i) Contour plot of the distribution function, t=1900,
(j) Contour plot of the distribution function, t=1920,
(k) Contour plot of the distribution function, t=1940,
(l) Contour plot of the distribution function, t=1960.



(m) (n)



(o)

Fig. 7. (m) Contour plot of the distribution function, t=2000,
(n) Contour plot of the distribution function, t=2200,
(o) Contour plot of the distribution function, t=2980.

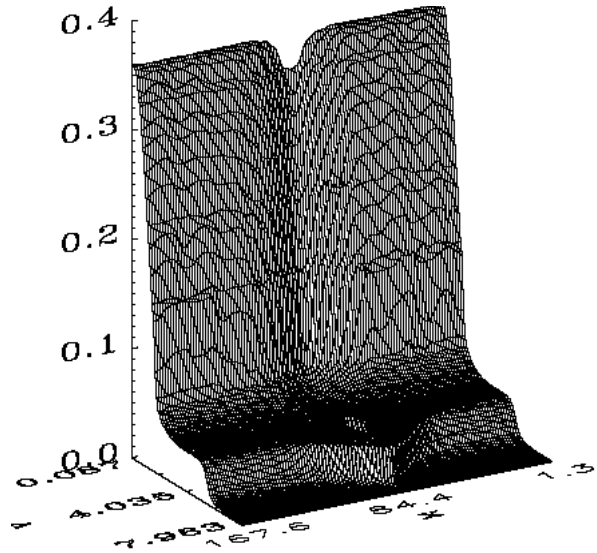


Fig. 8. Three-dimensional view of the results in Fig.7o.

distribution function in the region of the bulk, showing a small knee around $\upsilon \approx 1.1$ and around $\upsilon \approx 1.3$. This corresponds to longitudinal modulations we see in Fig.(8). Fig.(11c) shows, on a linear scale, the top of the electron distribution function, which shows a small cavity around $\upsilon \approx -0.05$. The acoustic speed in our normalized units is $\sqrt{T_e / M_i} / \sqrt{T_e / M_e} = \sqrt{M_e / M_i} \approx 0.023$. In the phase-space plot in Fig.(11d) the structure around $\upsilon \approx -0.05$ shows six vortices, corresponding to a mode with $k = 0.225$.

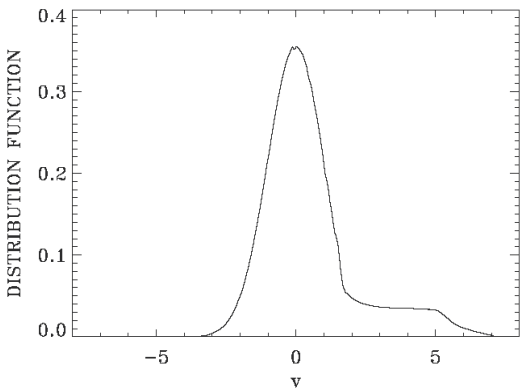


Fig. 9. Electron distribution function at $t = 2980$.

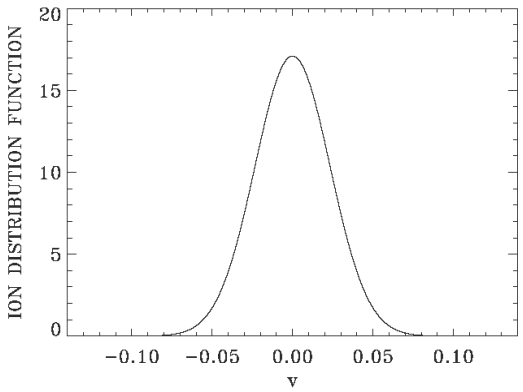


Fig. 10. Ion distribution function at $t = 2980$

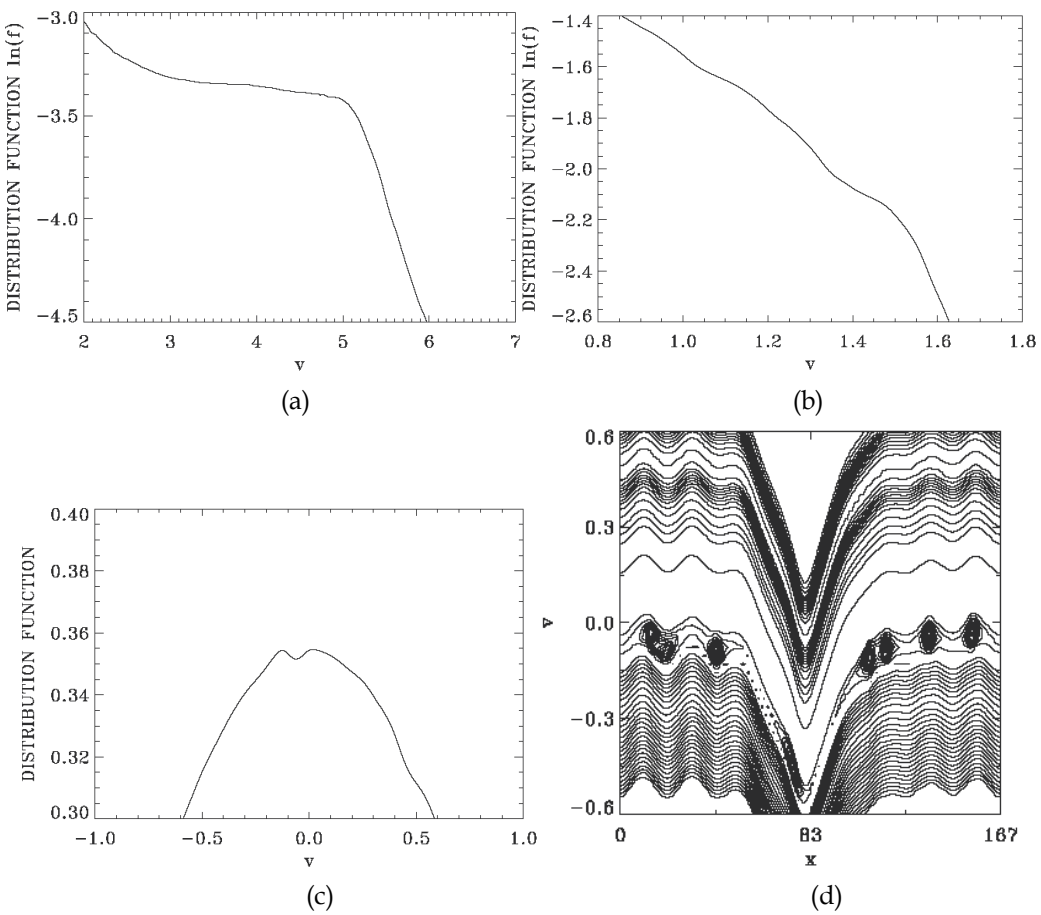
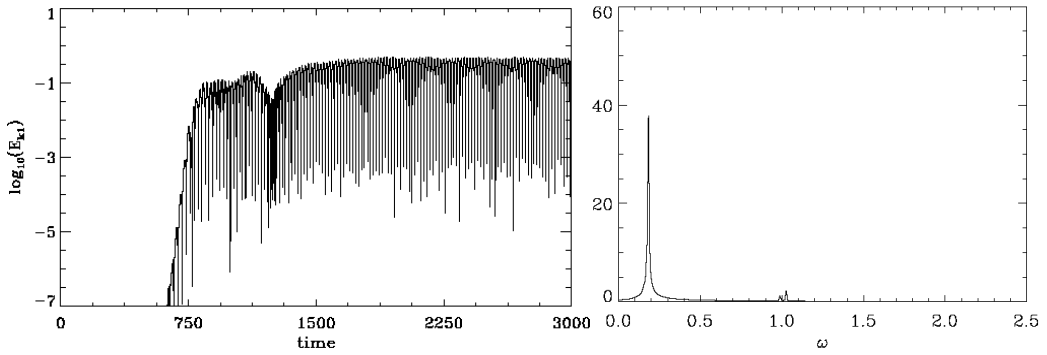


Fig. 11. (a) Same as Fig.(9) (concentrates on the tail)
(b) Same as Fig.(9) (concentrates on the bulk)
(c) Same as Fig.(9) (concentrates on the top)
(d) Contour plot for the distribution in Fig.(11c)

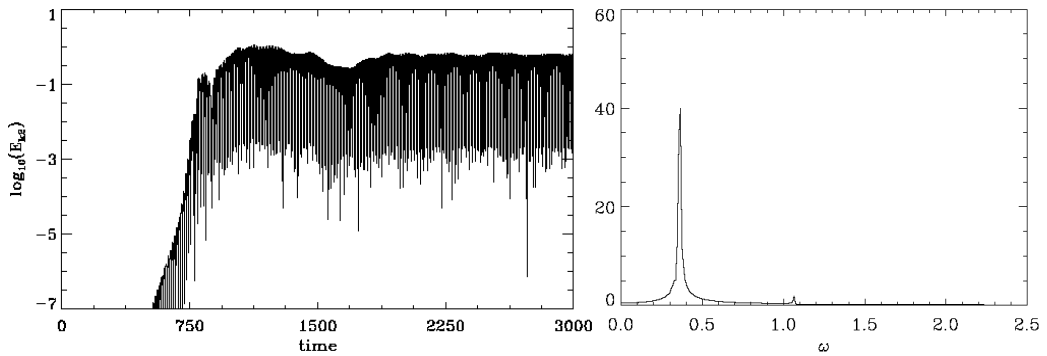
Figs.(12a-19a,20-22) show the time evolution of the different Fourier modes $k = n2\pi / L$ with n =1,2,3,4,5,6,7,8,9,12,16. Fig.(19) shows the initially unstable mode with $k = 0.3$, $n = 8$, growing then saturating (which corresponds to the eight vortices we see in Fig.(2)), and showing trapped particles oscillation. The merging of the vortices in the presence of growing sidebands for $t > 700$ is accompanied by an inverse cascade with a transfer of energy to longest wavelengths. We see the amplitude of the Fourier mode $k = 0.3$, $n = 8$ decreasing sharply for $t > 700$. Also the phase velocity of the center of the final hole in Fig.(7o) for instance has moved higher and is about 4.8, due to the acceleration of the particles during the merging of the vortices. The frequencies of these longest wavelengths are below the plasma frequency. We calculate the frequencies of the different modes by their Fourier transform in the steady state at the end of the evolution, from $t_1 = 2344$ to $t_2 = 3000$. The frequency spectrum of the mode $k = 0.0375$, $n = 1$ given in Fig.(12a) is shown in Fig.(12b), with a peak at $\omega = 0.182$, which corresponds to a phase velocity $\omega / k = 4.853$ around the center of the vortex in Fig.(7o). We have also in Fig.(12b) two very small peaks at $\omega = 0.9875$ and $\omega = 1.0258$, which are modulating the amplitude of the mode. The frequency spectrum of the mode $k = 0.075$, $n = 2$ in Fig.(13a) is given in Fig.(13b), which shows a peak at $\omega = 0.3643$, corresponding to a phase velocity $\omega / k = 4.857$. Another small peak is appearing at $\omega = 1.064$. The frequency spectrum of the mode $k = 0.1125$, $n = 3$ in Fig.(14a) is given in Fig.(14b), which shows a peak at $\omega = 0.5369$, corresponding to a phase velocity $\omega / k = 4.78$. The frequency spectrum of the mode $k = 0.15$, $n = 4$ in Fig.(15a) is given in Fig.(15b), which shows a peak at $\omega = 0.719$, corresponding to a phase velocity $\omega / k = 4.793$. The frequency spectrum of the mode $k = 0.1875$, $n = 5$ in Fig.(16a) is given in Fig.(16b), which shows a peak at $\omega = 0.901$, corresponding to a phase velocity $\omega / k = 4.805$. The frequency spectrum of the mode $k = 0.225$, $n = 6$ in Fig.(17a) is given in Fig.(17b), which shows a peak at $\omega = 1.0833$, corresponding to a phase velocity $\omega / k = 4.814$. The frequency spectrum of the mode $k = 0.2625$, $n = 7$ in Fig.(18a) is given in Fig.(18b), which shows a peak at $\omega = 1.0546$, and at $\omega = 1.256$, whose phase velocities are $\omega / k = 3.63$ and $\omega / k = 4.784$ respectively, corresponding to the two inflexion points we see in Fig.(11a) around $\upsilon \approx 3.63$ and $\upsilon \approx 4.8$. The frequency spectrum of the mode $k = 0.3$, $n = 8$ in Fig.(19a) is given in Fig.(19b) during the growth of the mode from $t_1 = 100$ to $t_2 = 755$, and in Fig.(19c) at the end from $t_1 = 2344$ to $t_2 = 3000$. During the first phase of the evolution of the mode in Fig.(19b) the dominant peak is at $\omega = 1.0258$ (reaching a peak of about 500), and other peaks are seen at $\omega = 0.7382$, 1.112, 1.313, 1.7928. For the steady state spectrum in Fig.(19c), the two dominant peaks are at $\omega = 1.1025$ and $\omega = 1.438$, whose phase velocities ω / k are respectively at 3.675 and 4.793, corresponding to the two inflexion points we see in Fig.(11a). We present in Figs.(20-22) the time evolution of the modes with $k = 0.3375$, $n = 9$, $k = 0.45$, $n = 12$ and $k = 0.6$, $n = 16$ (this last one is the harmonic of the mode $n = 8$ in Fig.(19)).

Figs.(23a,b) and Fig.(24) show respectively the electric field plot, the potential plot and the electron density plot at *t*=2980. Note the rapid variation of the electric field plot at the position of the hole in the phase-space in Fig.(7o), and the corresponding peak in the potential in Fig.(23b). Note the cavity-like structure at the position of the phase-space hole in the electron density plot in Fig.(24). The ions remained essentially immobile, and showed some effects during the evolution of the system, immobilizing a very small oscillation which



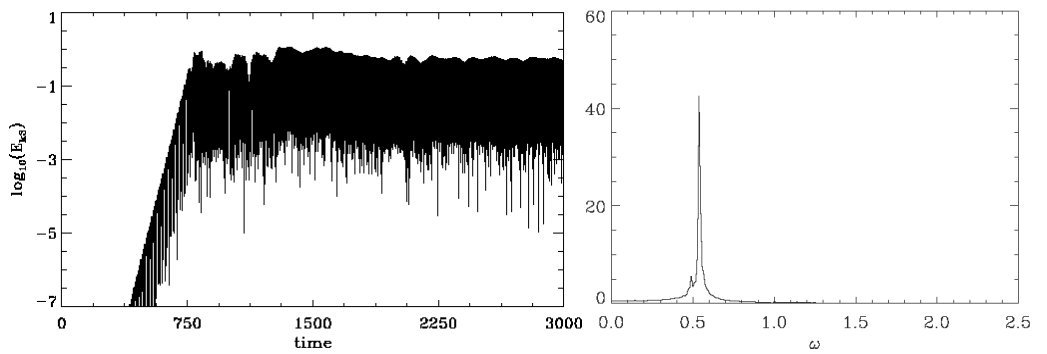
(a) (b)

Fig.12. (a) Time evolution of the Fourier mode k=0.0375
(b) Spectrum of the Fourier mode k=0.0375



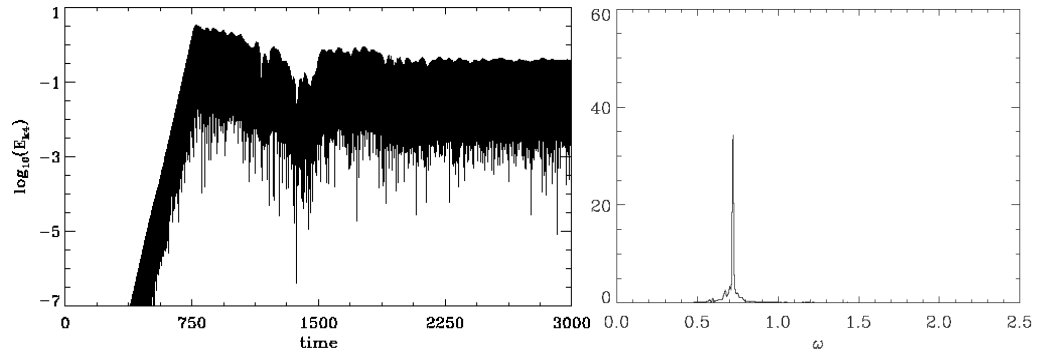
(a) (b)

Fig. 13. (a) Time evolution of the Fourier mode k=0.075
(b) Spectrum of the Fourier mode k=0.075



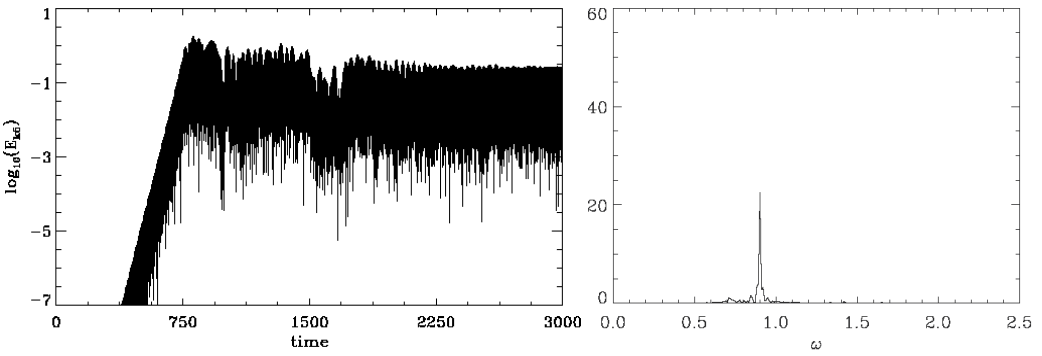
(a) (b)

Fig.14. (a) Time evolution of the Fourier mode with k=0.1125
(b) Spectrum of the Fourier mode k=0.112



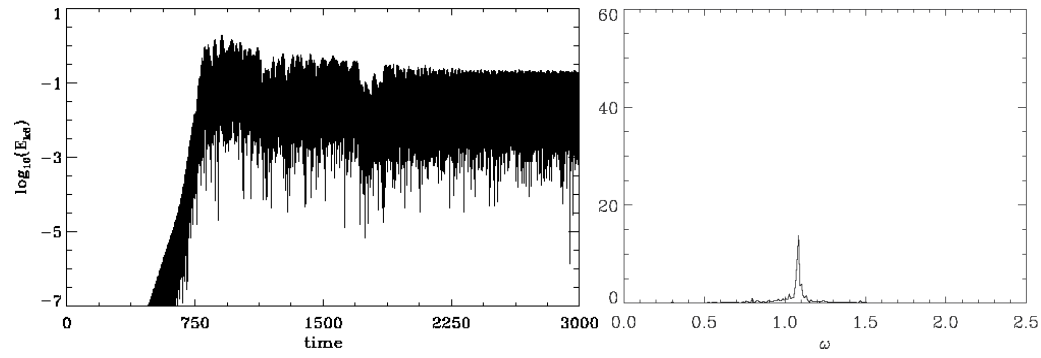
(a) (b)

Fig. 15. (a) Time evolution of the Fourier mode with k=0.15
(b) Spectrum of the Fourier mode k=0.15



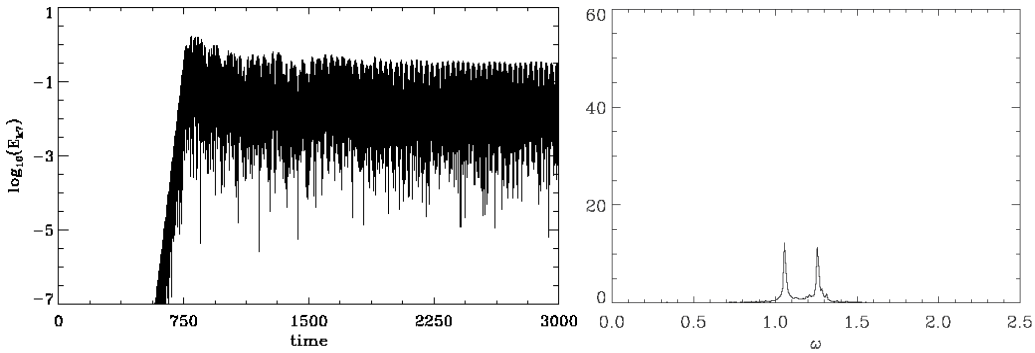
(a) (b)

Fig. 16. (a) Time evolution of the Fourier mode with k=0.1875,
(b) Spectrum of the Fourier mode k=0.1875



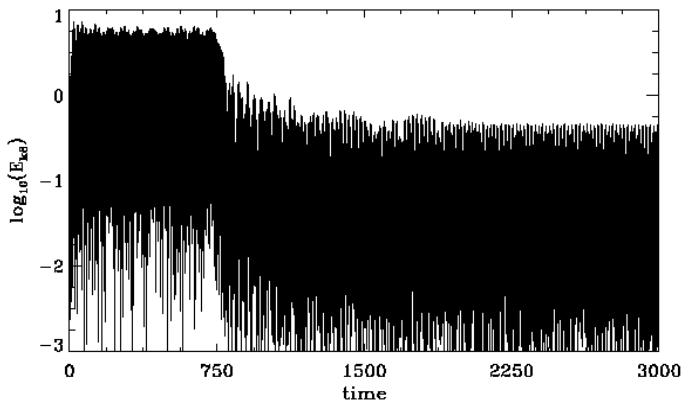
(a) (b)

Fig. 17. (a) Time evolution of the Fourier mode with k=0.225,
(b) Spectrum of the Fourier mode k=0.225

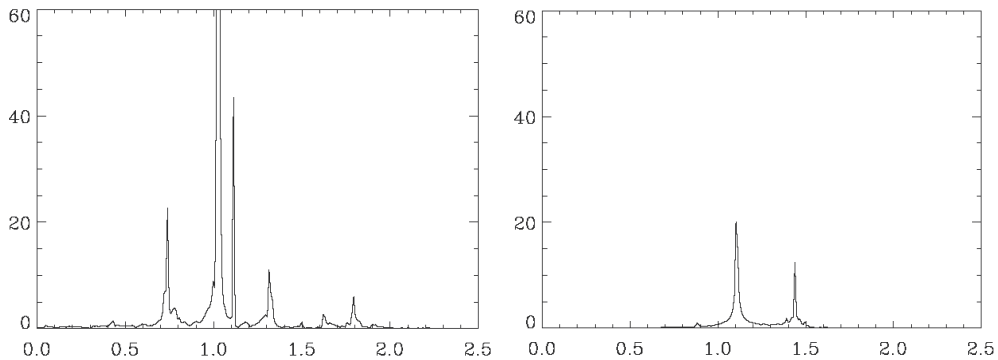


(a) (b)

Fig. 18. (a) Time evolution of the Fourier mode with k=0.2625,
(b) Spectrum of the Fourier mode k=0.2625



(a)



(b) (c)

Fig. 19. (a) Time evolution of the Fourier mode with k=0.3,
(b) Spectrum of the Fourier mode k=0.3 (from t=100. to t=755.36),
(c) Spectrum of the Fourier mode k=0.3 (from t=2344 to t=3000)

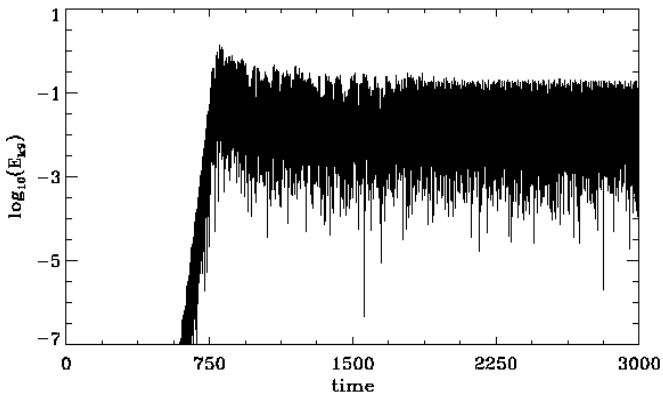


Fig. 20. Time evolution of the Fourier mode with k=0.3375

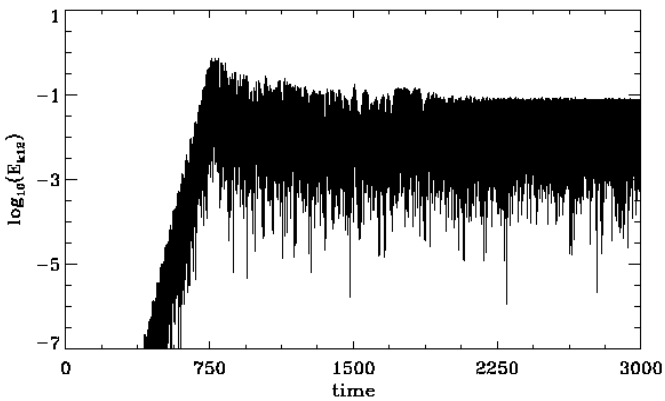


Fig. 21. Time evolution of the Fourier mode with k=0.45

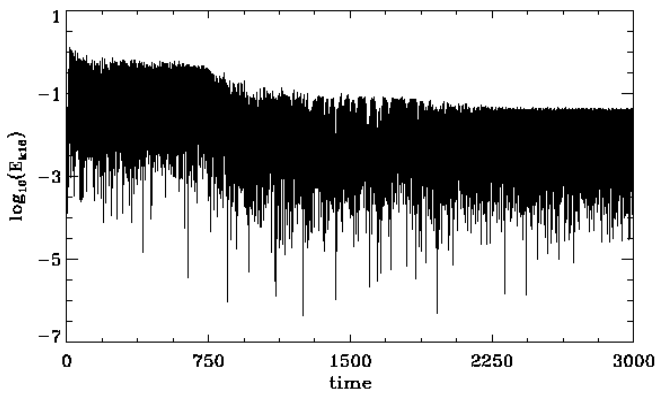
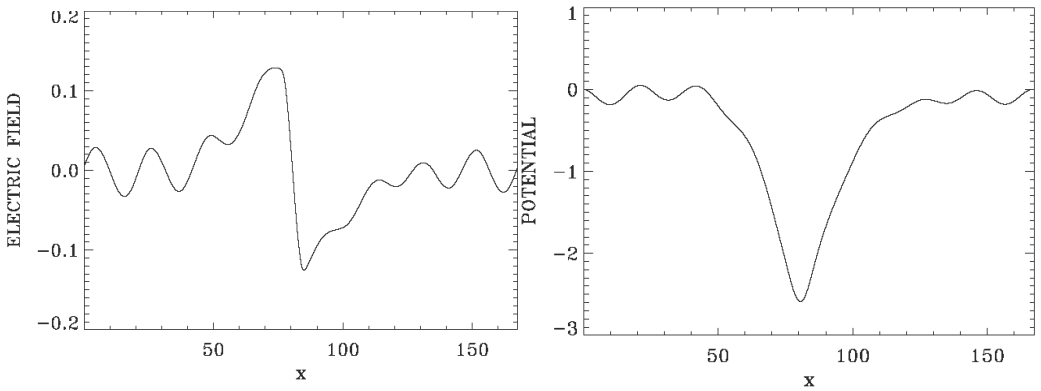


Fig. 22. Time evolution of the Fourier mode with k=0.6



(a) (b)

Fig. 23. (a) Electric field profile at *t*=2980 (b) Potential profile at *t*=2980.

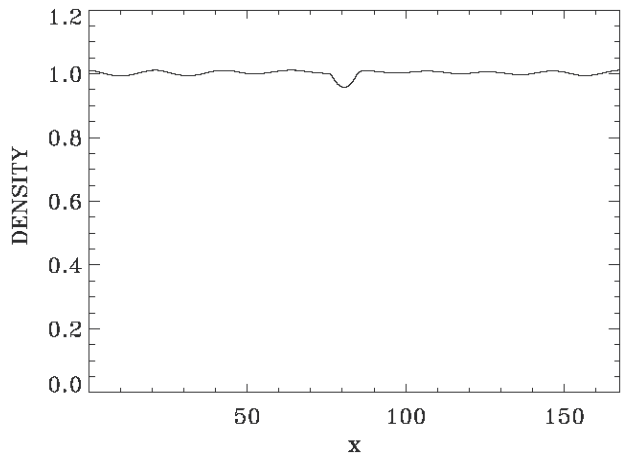


Fig. 24. Electron density profile at *t*=2980.

was persistent at the end of the simulation in the tail of the distribution function in Fig.(11a), without affecting the shape of the tail at all, especially what appeared to be the two inflexion points around $\upsilon \approx 3.7$ and $\upsilon \approx 4.8$. Also the evolution of the fusion of the two holes in Figs.(7h-7o) was much slower for the case of immobile ions, (lasting up to t=3000), with respect to what we see in the present results in Figs.(7h-7o) where the fusion is completed before *t*=2200.

4. Excitation of the modes n=7 and n=8 with k_a=0.2625 and k=0.3 respectively

We consider in this section the case when we excite initially two initially unstable modes with $k = n\frac{2\pi}{L} = 0.3$, $n = 8$, and $k_a = 0.2625$, $n = 7$. So the initial electron distribution function is given by:

$$f_e(x,\upsilon_e) = f(\upsilon_e)(1 + \varepsilon\cos(kx) + \varepsilon_a\cos(k_a x))) \tag{7}$$

$f_e(\upsilon_e)$ is defined in Eq.(4). We use $\varepsilon = \varepsilon_a = 0.04$. The linear solution for the frequency associated with the mode $k = 0.3$ is $\omega^2 \approx 1 + 3k^2$, or $\omega \approx 1.127$, with a phase velocity of the wave $\omega / k \approx 3.756$. The linear solution for the frequency associated with the mode $k_a = 0.2625$ is 1.0985, with a phase velocity of the wave $1.0985 / 0.2625 = 4.184$. Both phase velocities fall on the positive slope of the bump-on-tail distribution function, as can be verified from Fig.(3). So both initially excited modes are unstable.

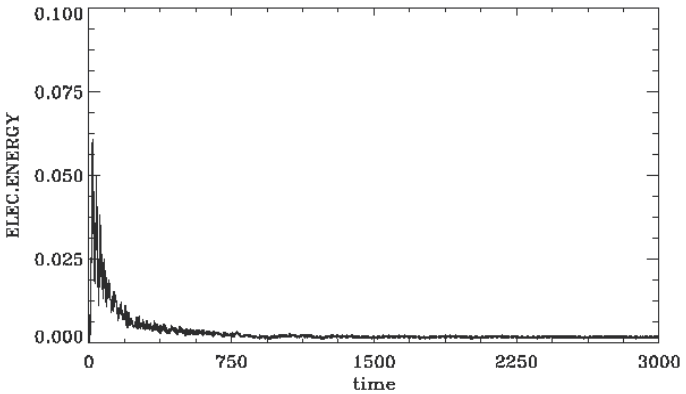


Fig. 25. Time evolution of the electric field energy.

Fig.(25) presents the time evolution of the electric field energy, which contrasts with what is presented in Fig.(1). Fig.(25) shows a rapid growth in the linear phase, followed by a rapid decay of the electric field energy. The spatially averaged electron distribution function shows very rapidly the formation of an elongated tail. We present in Fig.(26a) the spatially averaged electron distribution function at $t = 400$, and in Fig.(26b) we concentrate on the region of the tail, where the plot on a logarithmic scale show at this stage of the evolution a slowly decaying distribution function.

We present in Fig.(27a-o) the evolution of the phase-space. From the early beginning, the vortices formed due to the trapping of particles are unstable. Energy flows to the longest

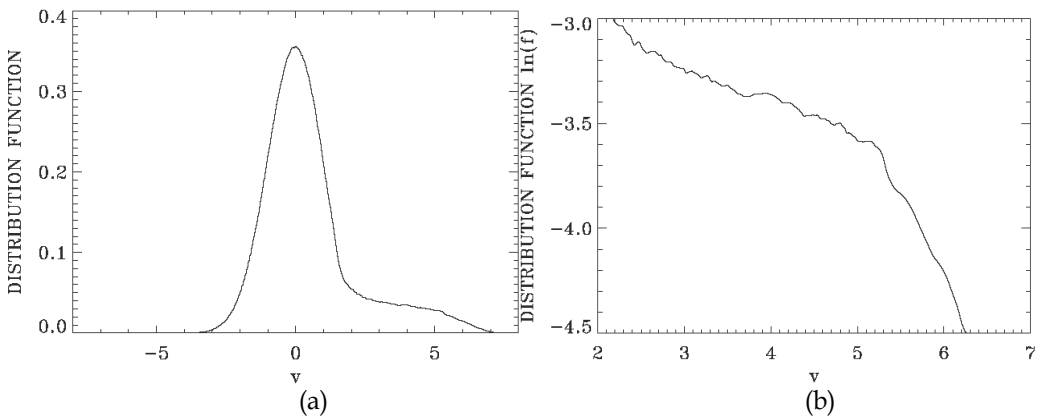


Fig. 26. (a) Spatially averaged distribution function at $t = 400$,
(b) Same as Fig.(26a), concentrating on the tail region.

wavelengths, which is characteristic of 2D systems (Knorr,1977). During this evolution the center of the vorticies is moving to higher velocities. In Fig.(27d) at $t = 600$, we have two holes left, which then start merging together. In Fig.(27f) at $t = 760$, one of the two vortices starts occupying a satellite position around the other, and then starts spiralling around it, leaving in the long run a single vortex (the evolution at this stage is similar to what has been presented in the previous section in Figs.(7i-n)). At $t = 3000$ in Fig.(27o), we show the final single vortex, centered around ≈ 5.05. Note also in Fig.(27g) the presence of a small vortex along the upper boundary. In Figs.(27i-j) this small vortex moves closer to the big vortex, and then in Figs.(27k-m) it goes spiraling around the big vortex. Fig.(28) is a 3D plot of the hole presented in Fig.(27o). Note the associated cavity structure in the bulk which travels as a solitary like structure in the phase-space. In Fig.(29a) we show the spatially averaged electron distribution function at $t = 3000$, and in Fig.(29b) we present on a logarithmic scale the same curve, concentrating in the tail region.

Although the initial evolution of the system is totally different from what we see in the previous section, the final result in Fig.(27o) showing a hole in the phase-space is close to what has been presented in the previous section. There are, however, important differences between the results in Fig.(27o) and the results in Fig.(7o). The hole in Fig.(27o) is centered at a higher velocity than the hole in Fig.(7o). We observe also the plot of the tail in Fig.(29b) being shifted to higher velocities than the plot of the tail in Fig.(11a). Indeed, in Fig.(29b) the inflexion points are around $\upsilon \approx 4.05$ and around $\upsilon \approx 5.05$, while in Fig.(11a) the inflexion points are around $\upsilon \approx 3.7$ and $\upsilon \approx 4.8$. We present in Fig.(30a) the same electron distribution function as in Fig.(29a) at $t = 3000$, concentrating at the top of the distribution function. There is a deformation at the top which appears more important than the one at the top of Fig.(11d). Also the contour plot in Fig.(30b) at the top of the electron distribution function shows a rich collection of small vortices, more important than what we observe in Fig.(11d). Fig.(31a) and Fig.(31b) present the electric field and the electron density profiles at $t = 3000$. See in Fig.(31a) the rapid variation of the electric field from a positive to negative value at the position of the phase-space hole in Fig.(27o). See in Fig.(31b) the cavity structure in the density plot at the position of the phase-space hole. The ions showed essentially very small variation, and a flat density profile. However, this small variation provides the stable profile in Fig.(29). In the absence of the ions, the profile in Fig.(29b) would show a very small oscillation.

Figs.(32-44) present the Fourier modes and their frequency spectra. We note from these figures that the initial growth of the longest wavelengths during the process of inverse cascade is higher with respect to what we see in Figs(12a-18a) for instance. There is a modulation in the asymptotic state which is more important in Figs.(32-44). The frequency spectrum is calculated by transforming the different Fourier modes in the last part of the simulation from $t_1 = 2344$ to $t_2 = 3000$. The frequency spectrum of the mode with $k = 0.0375$, $n = 1$ in Fig.(32a) shows a peak at $\omega = 0.19175$. The phase velocity of this mode $\omega / k = 5.11$. Two other small peaks appear in Fig.(32b) at $\omega = 0.9875$ and 1.0258. The frequency spectrum of the mode with $k = 0.075$, $n = 2$ in Fig.(33a) has a peak at $\omega = 0.374$, corresponding to a phase velocity ≈ 5. It has also two small peaks at a frequency $\omega = 0.9875$ and 1.0738. The frequency spectrum of the mode with $k = 0.1125$, $n = 3$ in Fig.(34a) has a peak at a frequency $\omega = 0.5656$ in Fig.(34b), corresponding to a phase velocity ≈ 5.03. It has also a small peak at $\omega = 0.997$. The frequency spectrum of the mode with $k = 0.15$, $n = 4$ in

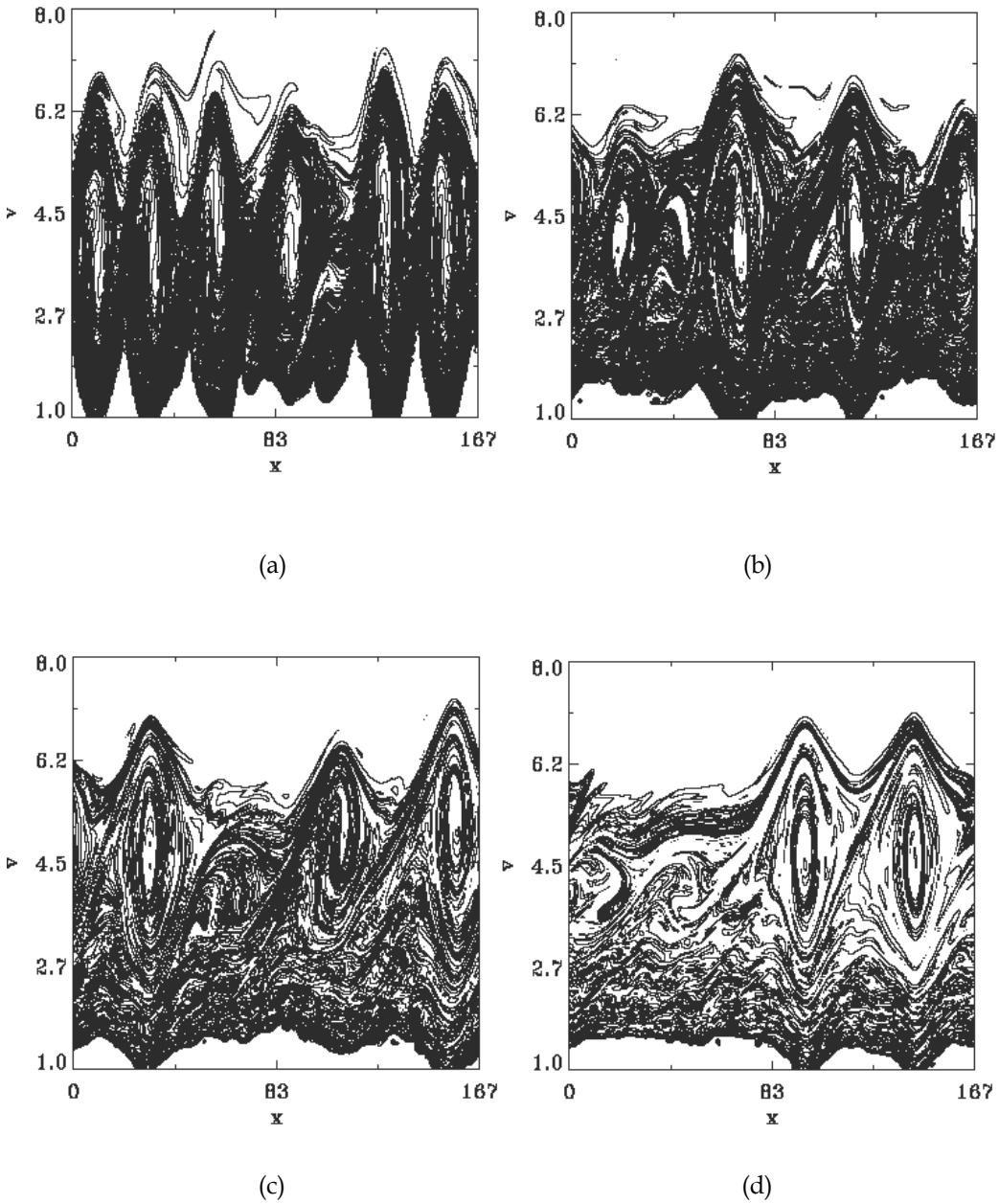


Fig. 27. (a) Contour plot of the distribution function, $t = 60$
(b) Contour plot of the distribution function, $t = 200$
(c) Contour plot of the distribution function, $t = 400$
(d) Contour plot of the distribution function, $t = 600$

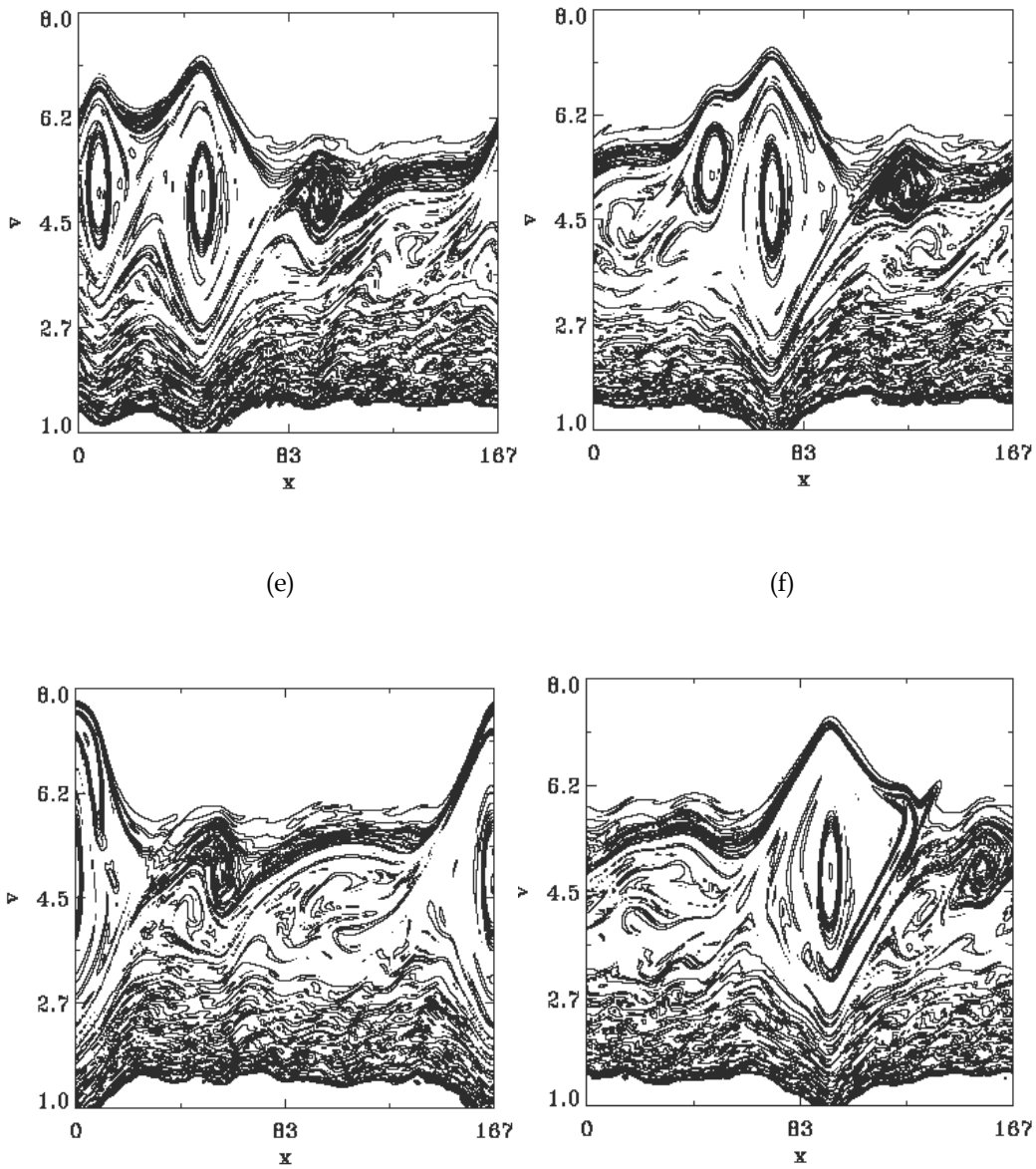


Fig. 27. (e) Contour plot of the distribution function, $t = 720$
(f) Contour plot of the distribution function, $t = 760$
(g) Contour plot of the distribution function, $t = 780$
(h) Contour plot of the distribution function, $t = 800$

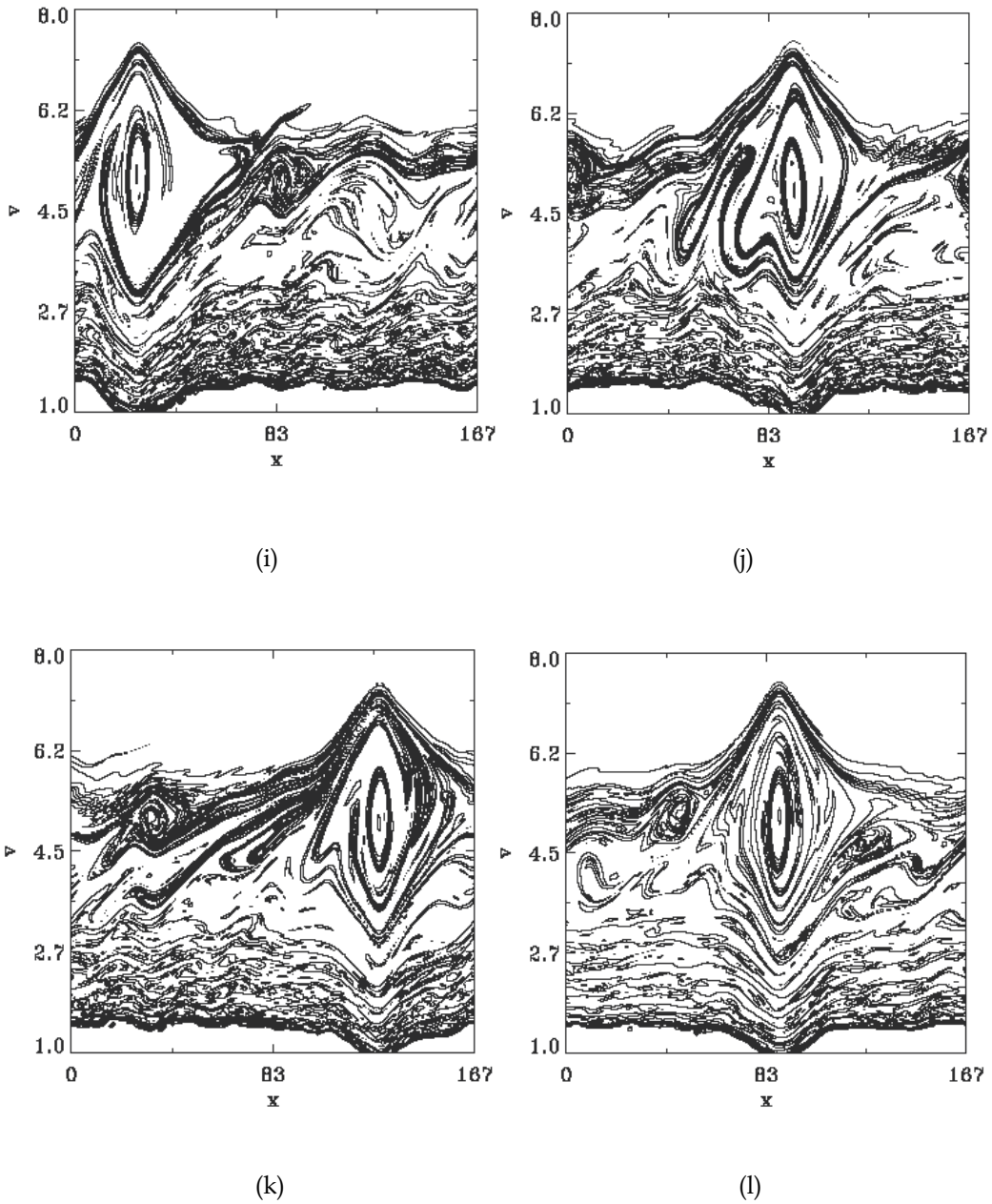


Fig. 27. (i) Contour plot of the distribution function, $t = 820$
(j) Contour plot of the distribution function, $t = 900$
(k) Contour plot of the distribution function, $t = 940$
(l) Contour plot of the distribution function, $t = 1200$

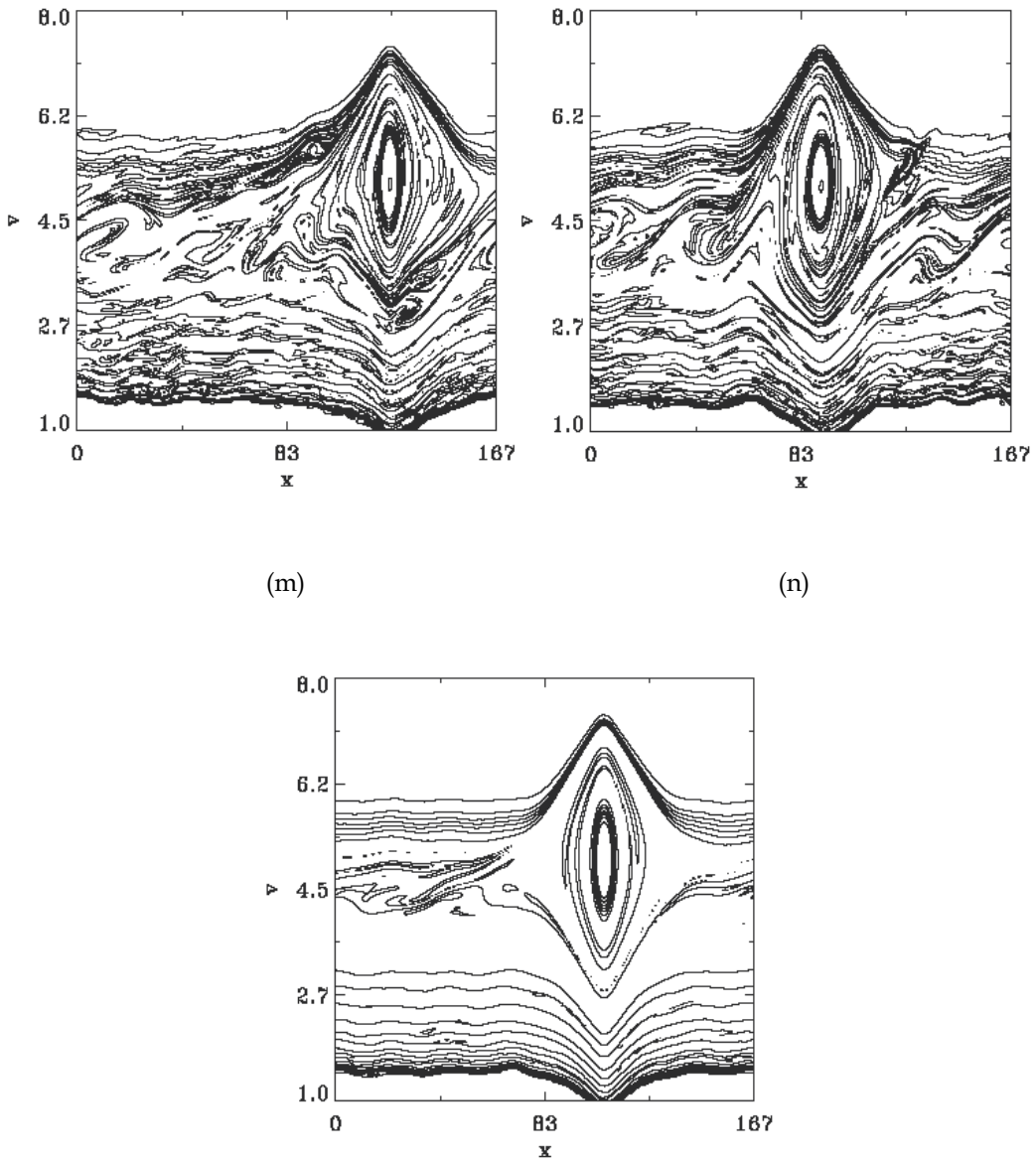


Fig. 27. (m) Contour plot of the distribution function, $t = 1240$
(n) Contour plot of the distribution function, $t = 1300$
(o) Contour plot of the distribution function, $t = 3000$

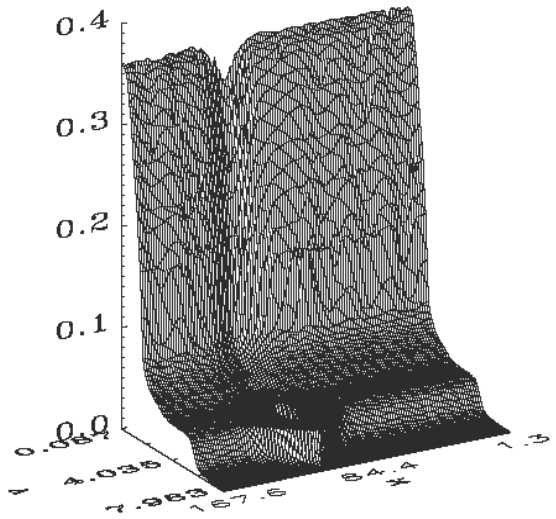
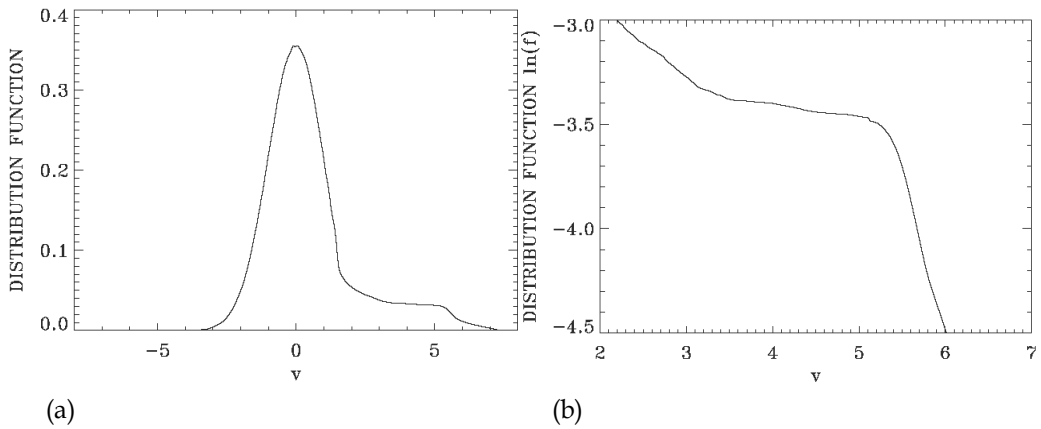


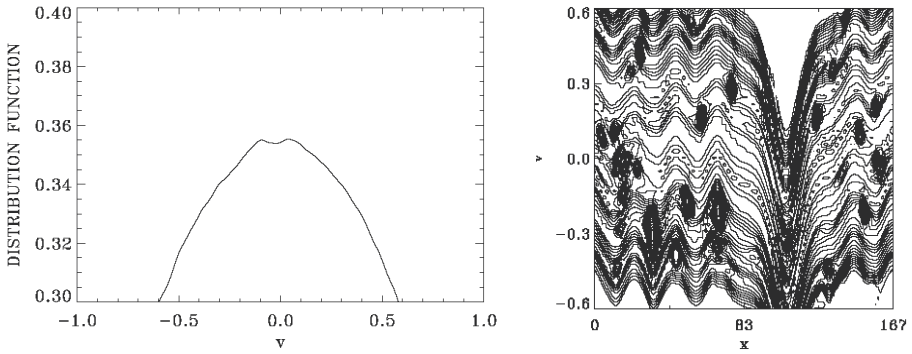
Fig. 28. Same as Fig.(27o), 3D plot at $t = 3000$



(a) (b)

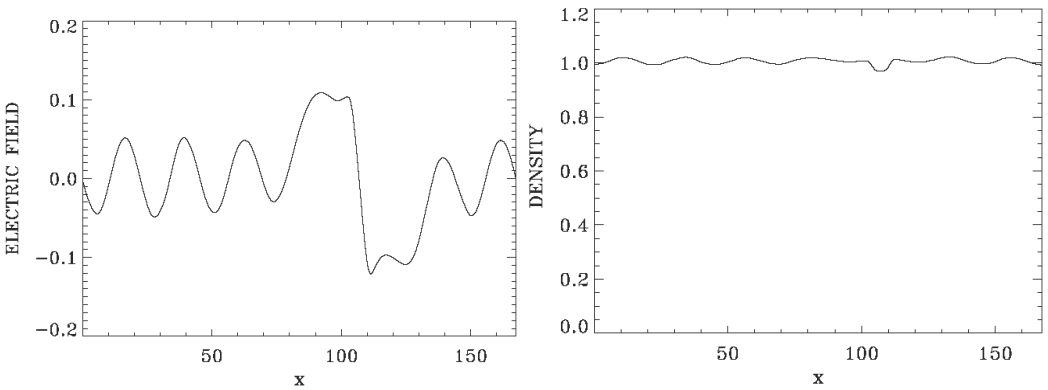
Fig. 29. (a) Spatially averaged distribution function, $t = 3000$
(b) Same as Fig.(29a), concentrating on the tail region.

Fig.(35a) has a peak at a frequency $\omega = 0.7574$ in Fig.(35b), corresponding to a phase velocity ≈ 5.05 . The frequency spectrum of the mode with $k = 0.1875$, $n = 5$ in Fig.(36a) has a peak at $\omega = 0.944$ in Fig.(36b), corresponding to a phase velocity ≈ 5.034 . The frequency spectrum of the mode with $k = 0.225$, $n = 6$ in Fig.(37a) has a peak at a frequency $\omega = 1.1313$ in Fig.(37b), corresponding to a phase velocity ≈ 5.028 . It has also peaks at $\omega = 1.0258$ and 1.256 , which underline the modulation of the mode in Fig.(37a). All the previous modes have a phase velocity ≈ 5.05 , which corresponds to the inflexion point of zero slope we see in Fig.(29b). So the dominant frequencies of oscillation of these modes seem to adjust themselves in such a way that the phase velocities of these modes would correspond to the



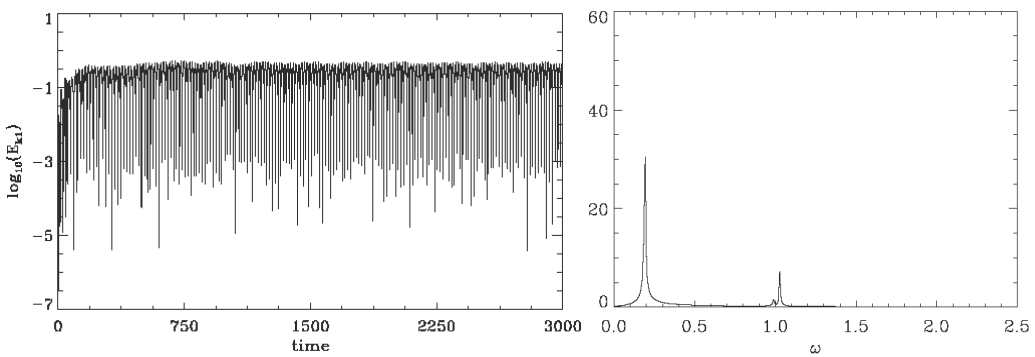
(a) (b)

Fig. 30. (a) Same as Fig.(29a) (concentrates on the top)
(b) Contour plot for the distribution in Fig.(30a)



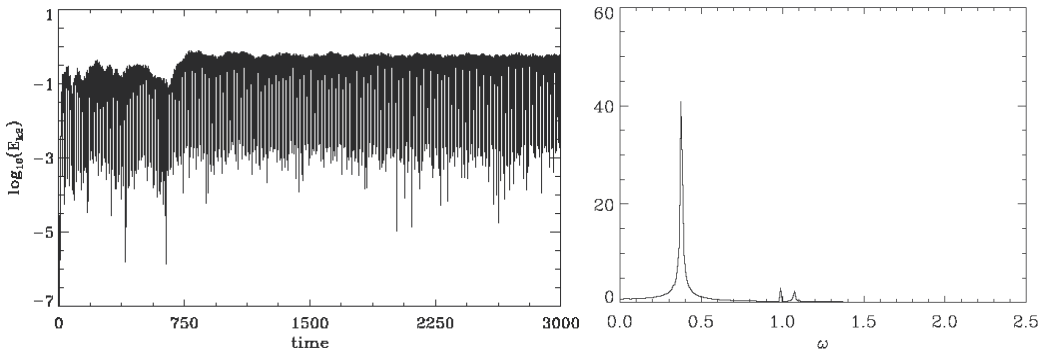
(a) (b)

Fig. 31. (a) Electric field profile at $t = 3000$,
(b) Electron density profile at $t = 3000$



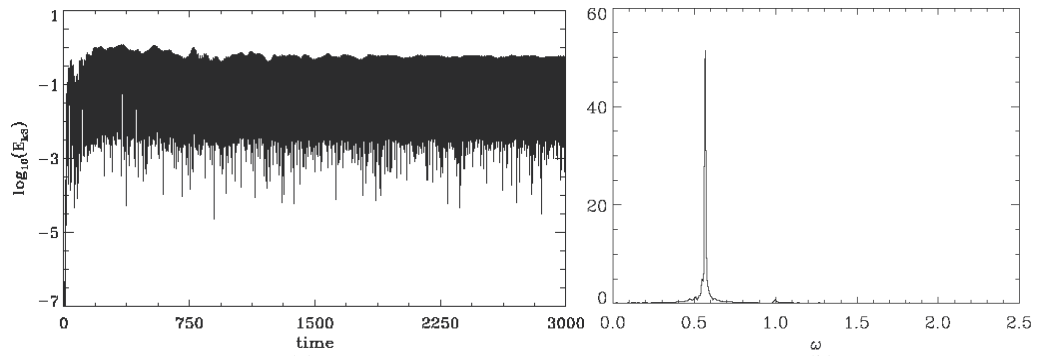
(a) (b)

Fig. 32. (a) Time evolution of the Fourier mode k=0.0375,
(b) Spectrum of the Fourier mode k=0.0375



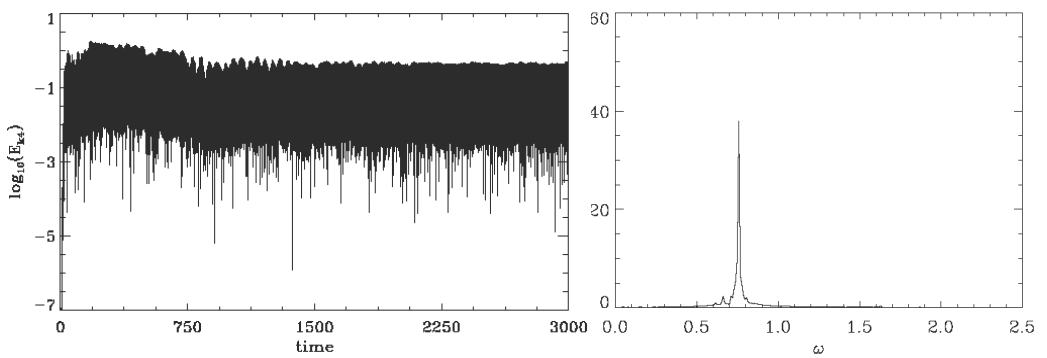
(a) (b)

Fig. 33. (a) Time evolution of the Fourier mode k=0.075,
(b) Spectrum of the Fourier mode k=0.075



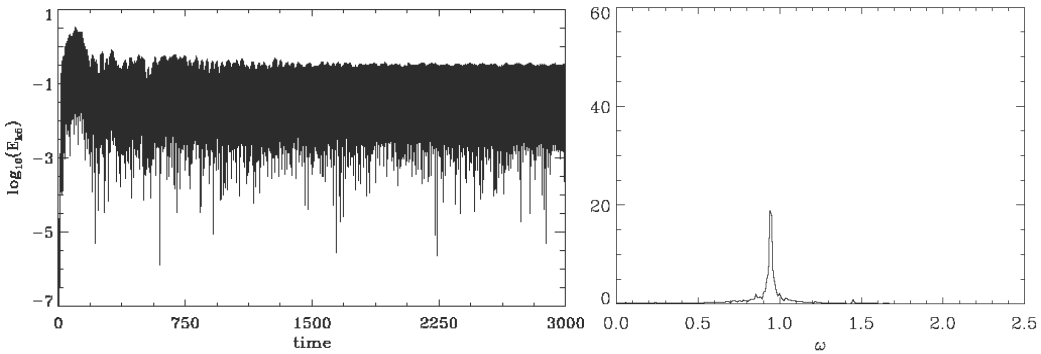
(a) (b)

Fig. 34. (a) Time evolution of the Fourier mode k=0.1125,
(b) Spectrum of the Fourier mode k=0.1125



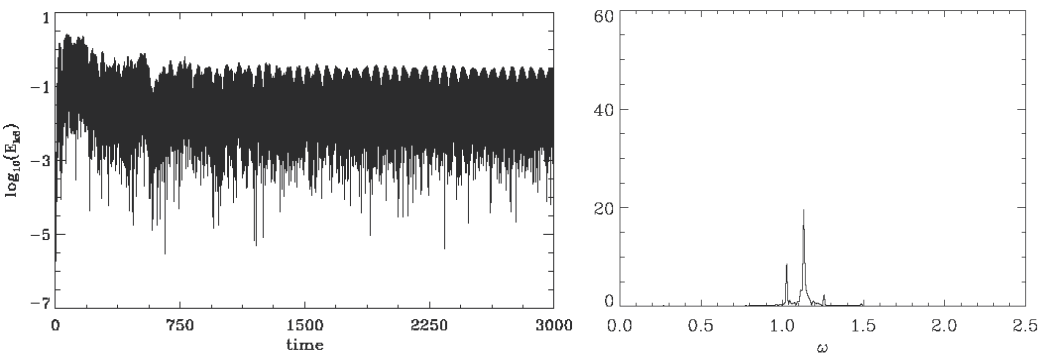
(a) (b)

Fig. 35. (a) Time evolution of the Fourier mode k=0.15
(b) Spectrum of the Fourier mode k=0.15



(a) (b)

Fig. 36. (a) Time evolution of the Fourier mode *k*=0.1875,
(b) Spectrum of the Fourier mode *k*=0.1875



(a) (b)

Fig. 37. (a) Time evolution of the Fourier mode *k*=0.225,
(b) Spectrum of the Fourier mode *k*=0.225

inflexion point established by the trapped population, to allow the mode to oscillate at a constant amplitude. The mode $n=7$ in Fig.(38a) is one of the two initially unstable modes. Fig.(38b) gives the frequency spectrum of this mode in the initial phase of the evolution from $t_1=100$ to $t_2=755$, showing a broad spectrum with two dominant peaks at $\omega=1.064$ and 1.227. In the steady state at the end of the simulation, the frequency spectrum of the mode with $k=0.2625$, $n=7$ has a dominant peak at a frequency $\omega=1.0642$ and a peak at $\omega=1.323$ in Fig.(38c), corresponding to a phase velocity respectively of ≈ 4.05 and ≈ 5.04. These two velocities correspond to the two inflexion points of zero slope we see in Fig.(29b). The mode $n=8$ in Fig.(39a) is also one of the two initially unstable modes. Fig.(39b) gives the frequency spectrum of this mode in the initial phase of the evolution, from $t_1=100$ to $t_2=755$, showing a broad spectrum with two dominant peaks at $\omega=1.112$ and 1.428. In the steady state at the end of the simulation, the frequency spectrum of the mode with $k=0.3$, $n=8$ has a peak at a frequency $\omega=1.112$ and a peak at a frequency 1.5148 in Fig.(39c), corresponding to a phase velocity respectively of ≈ 3.7 and ≈ 5.05. This second velocity corresponds to the inflexion point we see in Fig.(29b). The mode at $\omega=1.112$ would correspond to a coupling between the modes $n=1$ and $n=7$ (for the

wavenumbers $0.0375 + 0.2625 = 0.3$, and for the frequencies $0.182 + 1.0642 = 1.246$). The frequency spectrum of the mode with $k = 0.3375$, $n = 9$ in Fig.(40a) has a peak at a frequency $\omega = 1.6969$ in Fig.(40b), corresponding to a phase velocity ≈ 5.028, which corresponds to the inflexion point we see in Fig.(29b). Two small frequency peaks are also appearing at 0.6903 and 0.9204 and are also present in Fig.(40b). The frequency spectrum of the mode with $k = 0.375$, $n = 10$ in Fig.(41a) has a peak at a frequency $\omega = 1.8887$ in Fig.(41b), corresponding to a phase velocity ≈ 5.036, which corresponds to the inflexion point we see in Fig.(29b). The frequency spectrum of the mode with $k = 0.45$, $n = 12$ in Fig.(42a) has a peak at a frequency $\omega = 2.26262$ in Fig.(42b), corresponding to a phase velocity ≈ 5.028, which corresponds to the inflexion point we see in Fig.(29b). Finally Fig.(43) shows the time evolution of the mode with $k = 0.525$, $n = 14$ (the harmonic of the mode $n = 7$ in Fig.(38a)), and Fig.(44) shows the time evolution of the mode with $k = 0.6$, $n = 16$ (the harmonic of the mode $n = 8$ in Fig.(39a)).

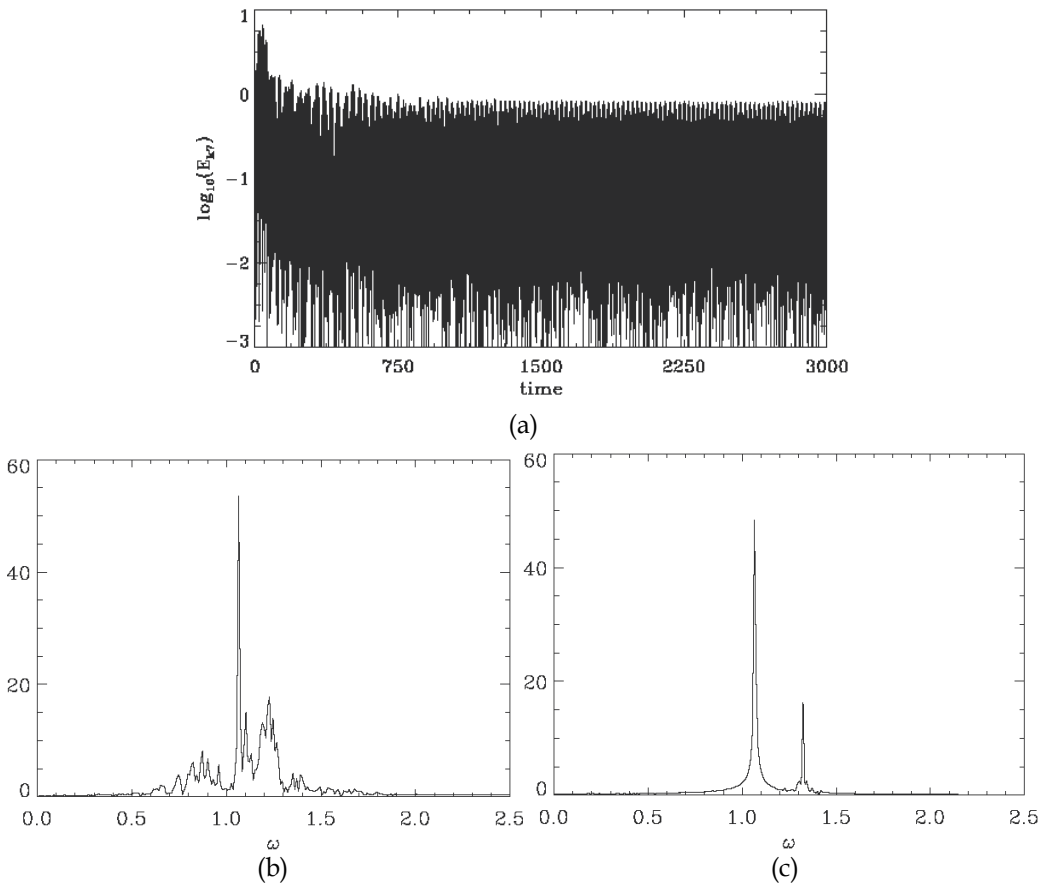


Fig. 38. (a) Time evolution of the Fourier mode k=0.2625,
(b) Spectrum of the Fourier mode k=0.2625 (from $t_1 = 100$ to $t_2 = 755$),
(c) Spectrum of the Fourier mode k=0.2625 (from $t_1 = 2344$ to $t_2 = 3000$)

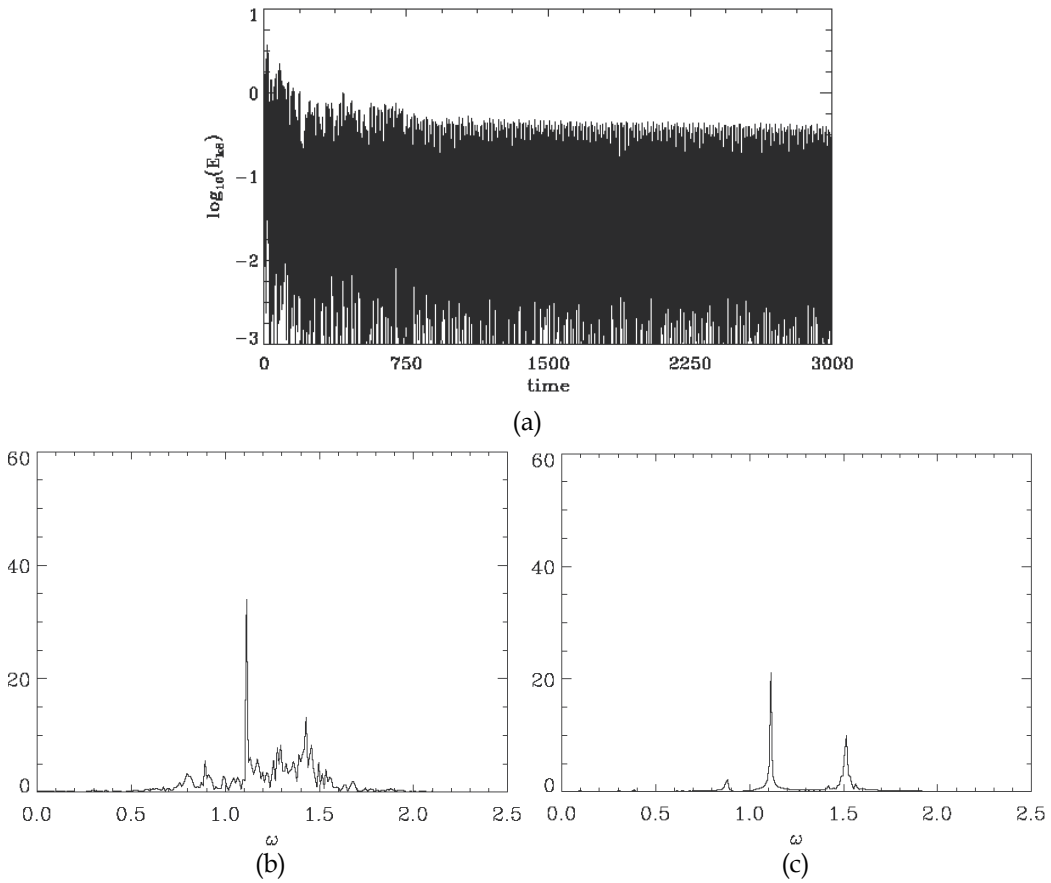


Fig. 39. (a) Time evolution of the Fourier mode k=0.3
(b) Spectrum of the Fourier mode k=0.3 (from $t_1 = 100$ to $t_2 = 755$
(c) Spectrum of the Fourier mode k=0.3 (from $t_1 = 2344$ to $t_2 = 3000$)

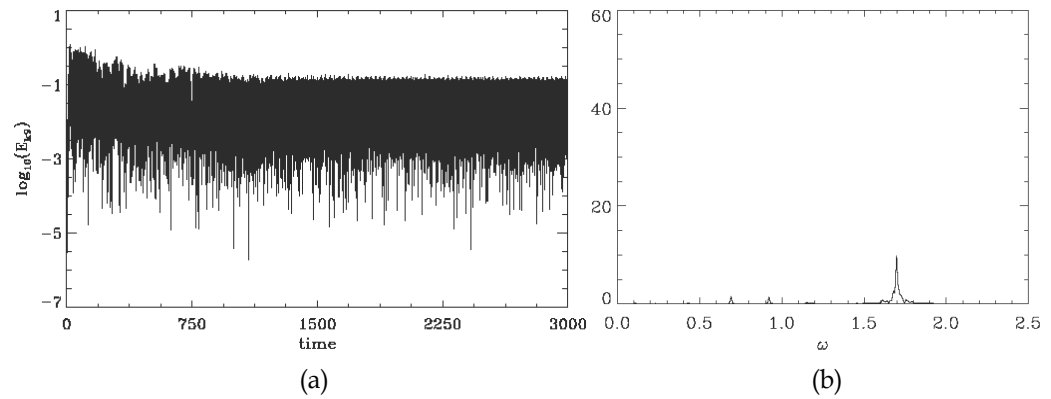
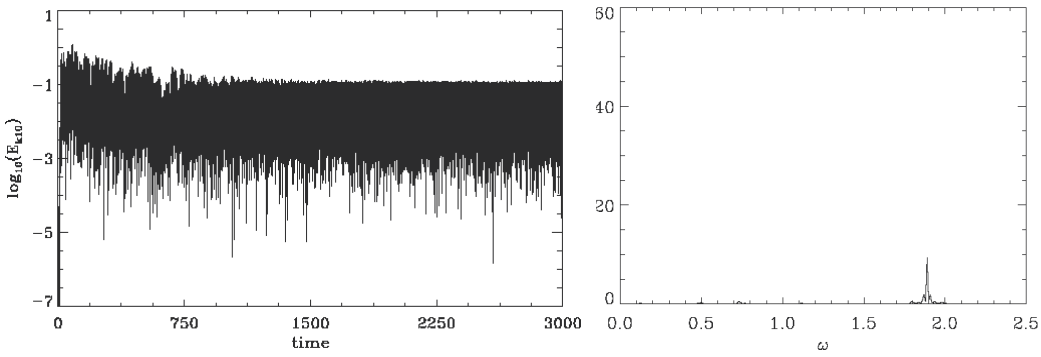
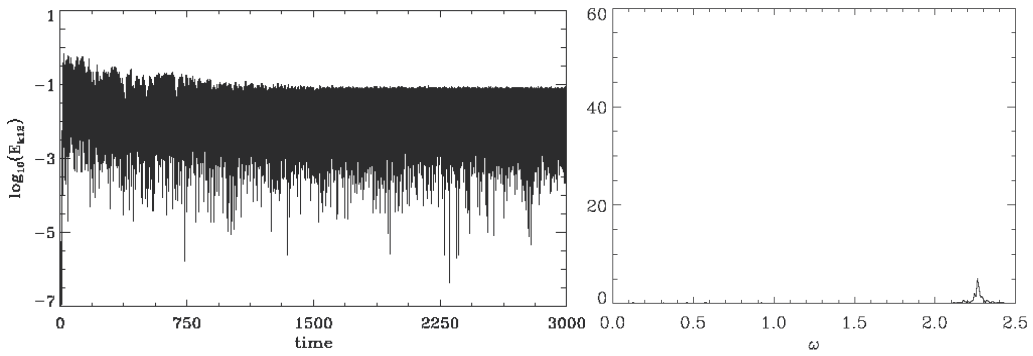


Fig. 40. (a) Time evolution of the Fourier mode k=0.3375,
(b) Spectrum of the Fourier mode k=0.3375



(a) (b)

Fig. 41. (a) Time evolution of the Fourier mode k=0.375,
(b) Spectrum of the Fourier mode k=0.375



(a) (b)

Fig. 42. (a) Time evolution of the Fourier mode k=0.45,
(b) Spectrum of the Fourier mode k=0.45

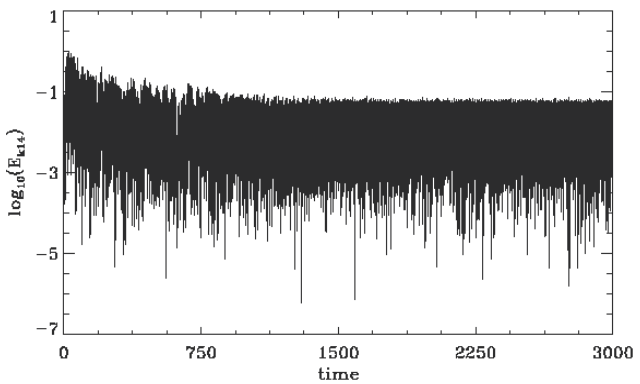


Fig. 43. Time evolution of the Fourier mode k=0.525

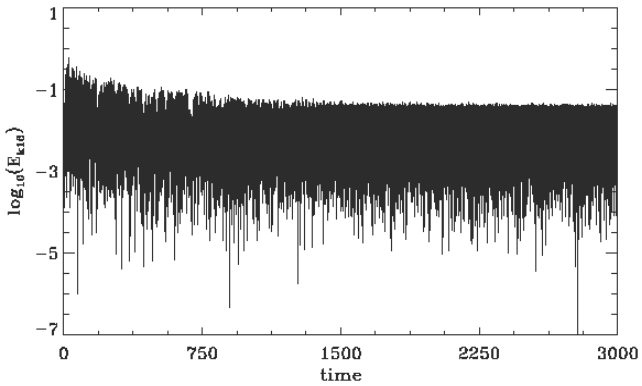


Fig. 44. Time evolution of the Fourier mode k=0.6

5. Conclusion

In the present Chapter, we have presented a study for the long-time evolution of the Vlasov-Poisson system for the case when the beam density is about 10% of the total density, which provides a vigorous beam-plasma interaction and an important wave-particle interaction, and which results in important trapped particles effects. A warm beam is considered, and the length of the system L is larger than the initially unstable wavelength λ, which allows the growth of sidebands. A fine resolution grid in the phase-space and a small time-step are used to follow the nonlinear dynamics of the trapped particles as accurately as possible. Numerical grid size effects and small time-steps can have important consequences on the number and distribution of the trapped particles, on the dynamical transitions of the Vlasov-Poisson system, and on kinetic microscopic processes such as the chaotic trajectories which appear in the resonance region at the separatrix of the vortex structures where particles can make transitions from trapped to untrapped and retrapped motion. The importance of the microscopic processes for their possible consequences on the macroscopic large scale dynamics have been also stressed in several publications (Califano *et al.*, 2000, Shoucri, 2010).

Initial conditions can also have important consequences on the microscopic processes, on the dynamical transitions of the Vlasov-Poisson system, and their possible effects on the large scale dynamics. Two cases have been considered in this Chapter. A case where a single initially unstable mode is excited, and a case where two initially unstable modes are excited. In the first case the evolution of the electric energy shows initially the classical behaviour of the growth of the initially unstable wave, followed by the saturation of the instability and the formation of BGK vortices in the phase-space, and the electric field is oscillating around a constant amplitude, modulated by the trapped particles oscillation. The subsequent evolution is dominated by a fusion of the vortices and by an inverse cascade where energy flows to the longest wavelengths available in the system, a process characteristic of 2D systems (Knorr, 1977). In the second case where two initially unstable waves are excited, the initial growth of the electric energy is followed by a rapid decay. This is accompanied by the formation of unstable vortex structures. In both cases the system evolves towards the formation of a single hole in the phase-space, where the trapped particles are accelerated to high velocities. The electron density plot shows the formation of a cavity like structure corresponding to the hole in the phase-space (Fig.(24) and (31b)).

Oscillation frequencies below the plasma frequency are associated with the longest wavelengths. The spatially averaged distribution functions show curves having a tail with a slowly decaying slope, and this slope takes the value of zero at the phase velocities of the dominant waves (see Fig.(11a) and (29b)). We note that for the second case in Fig.(29b), the acceleration of the particles in the tail is higher with respect to the first case in Fig.(11a). The low frequencies associated with the dominant longer wavelengths result in higher phase velocities of the different modes, which are accelerating trapped particles to higher velocities in the tail of the distribution function, with kinetic energies above the initial energy of the beam (see the recent work in Sircombe *et al*., 2006, 2008). The trapped accelerated population is adjusting in return in order to provide the distribution function with a zero slope at the phase velocities of the waves, allowing the different modes to oscillate at a constant amplitude (modulated by the oscillation of the trapped particles). The increase in the kinetic energy due to the particles acceleration is equivalent to the decrease in the electric energy we see in Fig.(1) and in Fig.(25).

The problem when several unstable modes are initially excited is certainly of interest. We note that when four initially unstable modes in a longer system were excited, the general evolution and the final results obtained were close to what we have presented in section 4 for the case of two unstable modes. Some additional results presented in Ghizzo *et al*. 1988, for this problem have shown a strong acceleration of particles in the case $n_b = 0.1$, in which case the tail particles are accelerated considerably to velocities higher than twice the initial beam velocity. The distribution function in this case takes the shape of a two-temperature Maxwellian distribution function with a high energy tail having a smooth negative slope. This result seems to agree with experimental observations from current drive experiments using an electron beam injected into the plasma (Advanced Concept Torus ACT-1 device), where it was observed that a significant fraction of the beam and background electrons are accelerated considerably beyond the initial beam velocity (Okuda *et al*., 1985). In none of the ACT-1 discharges is a distinctive feature of a plateau predicted from quasilinear theory apparent in the distribution function. The evolution of the waves amplitude in the results reported in Ghizzo *et al*., 1988, shows a rapid rise, followed by an abrupt collapse of the waves amplitude, the energy being delivered to the accelerated particles. When n_b is decreased, the acceleration of the particles is decreased, and when it is reduced to $n_b = 0.01$, a quasilinear plateau is formed and the waves amplitude saturate at a constant level. We note that when the simulations presented in sections 4 and 5 are repeated with $n_b = 0.01$, a quasilinear plateau is formed, without the acceleration we see in sections 4 and 5. Finally we point to the results obtained in Manfredi *et al*., 1996, with two spatial dimensions and a magnetic field, which shows in a bump-on-tail instability a rich variety of physics including also the acceleration of particles to high energies.

6. Acknowledgements

M. Shoucri is grateful to the Institut de Recherche d'Hydro-Québec (IREQ) computer center CASIR for computer time used to do the present work.

7. References

Berk, H.L., Roberts, K.V. (1967). Nonlinear study of Vlasov's equation for a special class of distribution functions. *Phys. Fluids* 10, 1595-1597

Bernstein, M., Greene, J.M., Kruskal, M.D. (1957). Exact nonlinear plasma oscillations. *Phys., Rev.* 108, 546-550

Bertrand, P., Ghizzo, A., Feix, M., Fijalkow, E., Mineau, P., Suh, N.D., Shoucri, M., (1988). Computer simulations of phase-space hole dynamics, In: *Nonlinear Phenomena in Vlasov Plasmas*, F. Doveil, (Ed.), p.109-125, Editions de Physique, Orsay.

Buchanan, M., Dorning, J. (1995). Nonlinear electrostatic waves in collisionless plasmas. *Phys. Rev. E* 52, 3015-3033

Califano, F., Lantano, M. (1999). Vlasov-Poisson simulations of strong wave-plasma interaction in condition of relevance for radio frequency plasma heating. *Phys. Rev. Lett.* 1999, 83, 96-99

Califano, F., Pegoraro, F., Bulanov, S.V. (2000). Impact of kinetic processes on the macroscopic nonlinear evolution of the electromagnetic-beam-plasma instability. *Phys. Rev. Lett.* 84, 3602-3605

Cheng, C.Z., Knorr, G. (1976). The integration of the Vlasov equation in configuration space. *J. Comp. Phys.* 22, 330-351

Crouseilles, N., Respaud, T., Sonnendrücker, E. (2009). A forward semi-Lagrangian method for the numerical solution of the Vlasov equation. *Comp. Phys. Comm.* 2009, 180, 1730-1745

Dawson, J.M., Shanny, R. (1968). Some investigations of Nonlinear behaviour in one-dimensional plasmas.*Phys. Fluids* 11, 1506-1523

Denavit, J., Kruer, W.L. (1971). Comparison of Numerical Solutions of the Vlasov equation with particle simulations of collisionles Plasmas. *Phys. Fluids* 14, 1782-1791

Doveil, F., Firpo, M.-C., Elskens, Y., Guyomarc'h, D., Poleni, M., Bertrand, P (2001). Trapping oscillations, discrete particle effects and kinetic theory of collisionless plasma. *Phys. Lett. A* 284, 279-285

Eliasson, B., Shukla, P.K. (2006). Formation and dynamics of coherent structures involving phase-space vortices in plasmas. *Phys. Rep.* 422, 225-290

Gagné, R., Shoucri, M. (1977). A splitting scheme for the numerical solution of the Vlasov equation. *J. Comp. Phys.* 24,445-449

Ghizzo, A., Shoucri, M.M., Bertrand, P., Feix, M., Fijalkow, E. (1988). Nonlinear evolution of the beam-plasma instabilities. *Phys. Lett.* A, 129, 453-458

Joyce, G., Knorr, G., Burns, T. (1971). Nonlinear behavior of the one-dimensional weak beam plasma system. *Phys. Fluids* 14, 797-801

Knorr, G. (1977). Two-dimensional turbulence of electrostatic Vlasov plasmas. *Plasma Phys.* 19, 529-538

Manfredi, M., Shoucri, M., Shkarofsky, I., Ghizzo, A., Bertrand, P., Fijalkow, E., Feix, M., Karttunen, S., Pattikangas, T., Salomaa, R. (1996). Collisionless diffusion of particles and current across a magnetic field in beam/plasma interaction. *Fusion Tech.* 29, 244-260

Nakamura, T.,Yabe, T. (1999). Cubic interpolated propagation scheme for solving the hyper-dimensional Vlasov-Poisson equation in phase-space. *Comput. Phys. Comm.* 120, 122-154

Nührenberg, J. (1971). A difference scheme for Vlasov's equation. *J. Appl. Math. Phys.* 22, 1057-1076

Okuda, H., Horton, R., Ono, M., Wong, K.L. (1985). Effects of beam plasma instability on current drive via injection of an electron beam into a torus. *Phys. Fluids* 28, 3365-3379

Pohn, E., Shoucri, M., Kamelander, G. (2005). Eulerian Vlasov codes. *Comm. Comp. Phys.* 166, 81-93

Schamel, H. (2000). Hole equilibria in Vlasov-Poisson systems: A challenge to wave theories of ideal plasmas.*Phys.Plasmas* 7, 4831- 4844

Shoucri, M. (1979). Nonlinear evolution of the bump-on-tail instability. *Phys. Fluids* 22, 2038-2039

Shoucri, M.(2008). *Numerical Solution of Hyperbolic Differential Equations*, Nova Science Publishers Inc.,New-York.

Shoucri, M.(2009). The application of the method of characteristics for the numerical solution of hyperbolic differential equations,In: *Numerical Solution of Hyperbolic Differential Equations*, S.P. Colombo,(Ed.), Nova Science Publishers, New-York.

Shoucri, M. (2010). The bump-on-tail instability, In: *Eulerian codes for the numerical solution of the kinetic equations of plasmas,* M. Shoucri, (Ed.), p. 291, Nova Science Publishers, New York.

Sircombe, N.J., Dieckman, M.E., Shukla, P.K., Arber, T.D. (2006). Stabilisation of BGK modes by relativistic effects. *Astron. Astrophys.452,* 371-382

Sircombe, N.J., Bingham, R., Sherlock, M., Mendonca, T., Norreys, P. (2008). Plasma heating by intense electron beams in fast ignition. *Plasma Phys. Control. Fusion* 50, 065005-(1-10)

Umeda, T., Omura, Y., Yoon, P.H., Gaelzer, R., Matsumoto, H. (2003). Harmonic Langmuir waves.III.Vlasov simulation. *Phys. Plasmas* 10, 382-391

Valentini, F., Carbone, V., Veltri, P., Mangeney, A. (2005). Wave-particle interaction and nonlinear Landau damping in collisionless electron plasmas. *Trans. Theory Stat. Phys.* 34, 89-101

2

Numerical Simulation of the Fast Processes in a Vacuum Electrical Discharge

I. V. Uimanov
Institute of Electrophysics of RAS
Russia

1. Introduction

A vacuum electrical discharge goes over three stages: breakdown, a spark, and an arc (Mesyats, 2000). The arc stage is the most mysterious form of the vacuum discharge. Since the discovery of the vacuum arc, the nature of the physical processes responsible for its operation has been debatable. The situation is paradoxical: vacuum arcs are widely used in various technologies, namely, in high-current switches, vacuum-arc melting and welding, plasma-assisted ion deposition, coating deposition, ion implantation, etc., and, at the same time, there is no commonly accepted idea about the mechanism of a vacuum discharge. Besides, the pulsed vacuum discharge is now the basic phenomenon harnessed in pulsed power and electronics. Pulsed electron beams are used in various fields, in particular for the production of braking x-ray pulses (radiography of fast processes, nondestructive testing), for the investigation of plasma–beam interactions, and for the production of superpower pulsed microwaves (heating of thermonuclear plasmas, long-range radar). It should be noted that practically in all cases, the spark (explosive electron emission (EEE)) stage of a vacuum discharge is used for the production of electron beams, and explosive-emission cathodes today have no alternative as components of pulsed high-current electrophysical devices. By the present time, the pulse range with the duration of processes as short as a few nanoseconds has been mastered rather well. Note that a way of increasing the power dissipated in the load of a high-voltage generator at a given stored energy is to reduce the voltage pulse duration. Therefore, in the recent years significant efforts have been made to develop high-current pulsed devices operating in the subnanosecond and picosecond ranges of voltage pulse durations (Mesyats & Yalandin, 2005).

It has been revealed in numerous experiments that the fundamental properties of a vacuum discharge are entirely determined by the processes that occur in a small brightly luminous region at the cathode through which the current transfer between the cathode and the electrode gap is realized. This region is termed the cathode spot and it includes the active part of the cathode surface heated to temperatures far above the melting point and the cathode plasma generated as a result of the material transfer from the active part of the cathode in the vacuum gap.

The existing theoretical models of the cathode spot phenomena can be conventionally subdivided into two groups in relation to the mechanism of formation of the conducting medium in vacuum. The models of the first group are based on the assumption of thermal

evaporation of the cathode material (Boxman et al., 1995; Lyubimov & Rakhovskii, 1978). However, this approach does not allow one to interpret well-known experimental data. The second-group models suppose explosive generation of the cathode plasma as a result of intense heating of microregions of the cathode surface. Theoretical ideas about explosive generation of the cathode plasma have been formulated most adequately in the ecton model of the cathode spot (Mesyats, 2000). In terms of this concept, all the three stages of a vacuum discharge appear explicable on the basis of natural origins. Breakdown and the so-called prebreakdown phenomena constitute a process of energy concentration in a microvolume at the cathode surface. Once the specific energy stored in a microvolume has become higher than a certain limiting value, an explosion begins and the breakdown stage is completed. The beginning of an explosion and the appearance of EEE is the onset of the spark stage of the discharge. The spark stage involves continuous regeneration of microexplosions by the plasma and liquid-metal jets produced by preceding microexplosions. The spark stage naturally goes over into the arc stage as the cathode and anode plasmas come together and the current rise rate decreases. However, the quantitative description of the cathode phenomena in the context of this model has yet been made only based on simplified estimating notions.

Thus, despite the significant advance in the study of some characteristic of cathode spots, a commonly accepted model of the cathode spot of a vacuum discharge yet does not exist. This is related, first, to the problems of experimental diagnostics of cathode spots in view of the extremely small time and space scales of the cathode spot phenomena and of their fast and chaotic motion over the cathode surface. Therefore, numerical simulation still remains practically a single method allowing one to determine the discharge parameters taking into account their space and time dependences.

The present paper is devoted to a numerical simulation of the prebreakdown phenomena in the cathode that take place on the initiation of an electrical discharge in a vacuum gap by application of a pulsed high voltage to the electrodes and also to a simulation of the processes initiating explosive emission centers on the cathode upon its interaction with the cathode spot plasma. Both processes in fact determine the mechanism of generation of the conducting medium in the electrode gap; therefore, to study them is of importance for constructing a self-consistent model of a vacuum discharge. In the first case, attention is mostly given to the subnanosecond range of voltage pulse durations. From the practical viewpoint, knowledge of the phenomena occurring on the nanosecond scale would be helpful in analyzing the efficiency of operation of explosive emission cathodes in a picosecond pulse mode. As for the second problem, the plasma–cathode interaction is the dominant process in the mechanism of self-sustainment of a vacuum discharge.

It is well known that under the conditions of high vacuum and pure electrodes, electrical discharge in vacuum is initiated by the current of field electron emission (FEE). According to the criterion for pulsed breakdown (Mesyats, 2000), to attain subnanosecond time delays to the explosion of a cathode microprotrusion, a field emission current density over 10^9 A/cm^2 is necessary. It is obvious that at these high field emission current densities, the screening of the electric field at the cathode surface by the space charge of emitted electrons substantially affects the field strength. It has even been speculated that this effect can have grave consequences: it will be impossible to produce current densities which would be high enough to achieve subnanosecond explosion delay times. The second section of proposed work is devoted to a study of the effect of the space charge of emitted electrons on the

electric field strength near the surface of the field emission emitters and a point microprotrusion on a metal cathode by using a two-dimensional axisymmetric problem statement. Based on the particle-in-cell (PIC) method, a model has been developed and self-consistent calculations of the electric field strength at the cathode and its field emission characteristics has been performed. In the third section a two-dimensional, two-temperature model has been developed to describe the prebreakdown phenomena in a cathode microprotrusion at picosecond and subnanosecond durations of the applied voltage pulse. The simulation procedure includes a particle-in-cell simulation to calculate the self-consistent electric field at the cathode surface and the field-emission characteristics of the cathode. In the fourth section a two-dimensional nonstationary model of the initiation of new explosive centers beneath the plasma of a vacuum arc cathode spot has been developed. In terms of this model, the plasma density and electron temperature that determine the ion current from the plasma to the microprotrusion and the microprotrusion geometry were treated as the external parameters of the problem. The process of heating of a cathode surface microprotrusion, for which both a surface irregularity resulting from the development of a preceding crater and the edge of an active crater, which may be a liquid-metal jet, can be considered, has been simulated numerically.

2. PIC simulation of the screening of the electric field at the cathode surface under intense field emission

The fact that the space charge (SC) of the electrons emitted from a metal affects the metal field- emission characteristics is now beyond question. The problem was first raised (Stern et al., 1929) shortly after the creation of the Fowler–Nordheim (F–N) theory (Fowler & Nordheim, 1928; Nordheim, 1929). However, it became urgent once Dyke and Trolan (Dyke & Trolan, 1953) had revealed an appreciable deviation from the F–N law at current densities $j > 5 \cdot 10^6$ A/cm^2, which showed up in a weaker dependence of the emission current on the applied potential difference. The authors (Dyke & Trolan, 1953) accounted for the nonlinearity of the current-voltage characteristics (CVCs) by the reduction of the electric field at the cathode surface due to the presence of a space charge of emitted electrons. In the subsequent work (Barbour et al., 1963), they proposed a model of an equivalent planar diode (EPD). This model, with properly chosen parameters, allowed one not only to describe qualitatively the deviation of a CVC toward the lower currents due to the SC effect, but also to obtain a reasonable agreement with experimental data. Therefore, for a long time it was considered established, both experimentally and theoretically, that the field-emission current density is limited to a level of ~10^7 A/cm^2 by the emission beam SC.

However, the problem appears to be not conclusively solved if field emission studies involve nanostructured surfaces and emitters where the emission occurs from nanometer objects. Investigations of the FEE from specially produced nanometer protrusions (Pavlov et al., 1975; Fursey et al., 1998) have shown that linear CVCs can be observed for current densities up to ~10^{10} A/cm^2, which are three orders of magnitude greater than those characteristic of conventional metal emitters with a tip radius of ~$10^{-5} \div 10^{-4}$ cm. Thus, though the current density is undoubtedly the determining quantity in the formation of the SC of a field emission beam, it is not the only factor responsible for the substantial effect of the SC on FEE. This necessitates a more rigorous consideration of this problem by invoking models that would describe the formation and spatial relaxation of the SC of an emission beam in a

more realistic geometry than this is possible in the context of the EPD model. This problem is also important in view of considerable advances in the study of field emission properties of various nanostructured surfaces, carbon nanotubes (Guillorn et al., 2004), dielectric and semiconductor matrices with conducting inclusions (Forbes, 2001), and Spindt cathodes (Spindt, 1968). In high current electronics, these investigations are of interest from the viewpoint of evaluating the efficiency of explosive-emission cathodes for the production of picosecond electron beams, because to initiate EEE within such short times, rather high FEE current densities (~10^{10} A/cm^2) are necessary (Mesyats & Uimanov, 2008).
Theoretically, the investigation of the effect of the SC on FEE was practically limited to the solution of the one-dimensional Poisson equation for an EPD or for a spherical diode (ESD) (see (Shrednik, 1974) and the cited literature). The one-dimensional approach naturally used in the previous work considerably moderates computational difficulties, but even in these cases, numerical calculations are required. This in the main is due to the nonlinearity of the F–N relation, which is used as a boundary condition in the problem statement. However, the applicability of a one-dimensional approximation to the actual geometry of a point-cathode vacuum diode has not been yet strictly substantiated. The only argument in favor of the usability of the EPD model advanced by the authors of Ref. (Barbour et al., 1963) is the estimate of the parameter of spatial localization of the SC near the emission surface. A critical analysis of the use of the EPD and ESD models for the description of the effect of the emission beam SC can be found elsewhere (Pavlov, 2004). The EPD model was adapted to describe the effect of the SC of emitted electrons on the field strength and current density distributions over the emitter surface (Shkuratov et al., 1960 (1995)). A similar approach was used with the ESD model (Batrakov et al., 1999). It has been shown (Shkuratov et al., 1960 (1995); Batrakov et al., 1999) that the SC of an emission beam not only efficiently screens the field at the cathode, but also significantly changes its distribution over the surface. However, it remains unclear for today whether the use of these quasi-two-dimensional approaches, offered largely *ad hoc*, is adequate. It should be noted that the particle-in-cell method was first used for solving the problem under consideration in Ref. (Batrakov et al., 1999). However, in our opinion, its capabilities, as applied to solving problems of this type, could not be efficiently used in spherical one-dimensional calculations. We were the first to make an attempt to solve the problem on the effect of the SC of emitted electrons on the electric field strength and on the CVC of the vacuum gap in a two-dimensional axially symmetric statement (Uimanov, 2008; Uimanov, 2010). We used the weighed-particle-in-cell method to simulate the self-consistent field-emission beam emitted by a microprotrusion on a macropoint cathode. The results obtained with the model developed have allowed us to analyze both the details of the screening phenomenon and the probable values of fields and current densities for the cathode protrusions of micrometer and submicrometer dimensions. In the study we present here, we used this model to investigate the external field screening not only for macropoint cathodes with microprotrusions, but also for classical point field emitters over a rather wide range of the geometric parameters of the cathode.

2.1 Problem statement and task geometry

Figure 1 presents the model geometry of the problem. As a whole it is the coaxial diode with distance the cathode - anode 1 cm. The cathode is the metal needle with the tip radius r_c.
On the surface of the cathode there is a microprotrusion of height h_m, tip radius r_m and the half-angle of the conical part Θ .

This cathode geometry takes into account the two-factor field enhancement at the microprotrusion surface which is typical of the electrode systems that were used in the experimental studies of EEE performed by now on the subnanosecond scale.

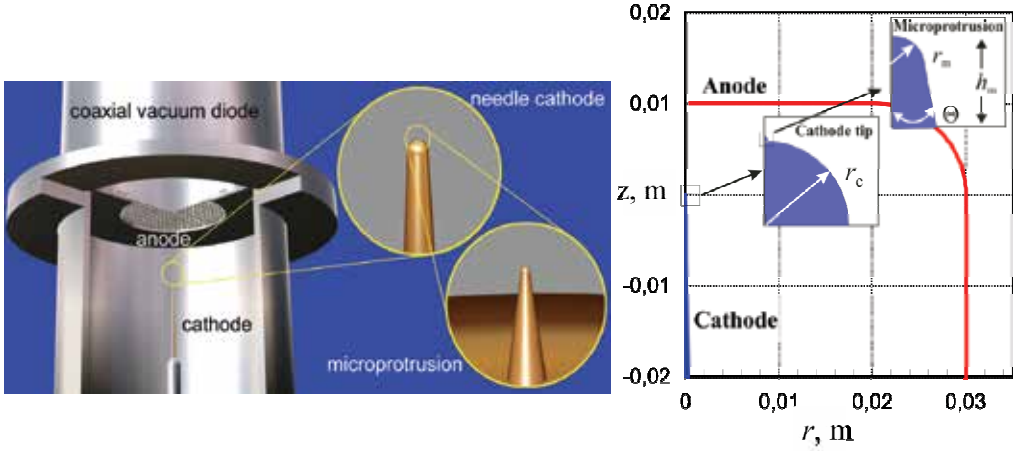


Fig. 1. Task geometry. Calculated parameters: r_c = 50 μm, h_m = 5 μm, r_m = 0.1 μm, $\Theta = 10^\circ$

Approximately the factor of electric field enhancement for such geometry is $\beta_{tot} = \beta_c \beta_m$, where β_c is the factor of electric field enhancement of the point cathode, β_m is the factor of the microprotrusion. The large difference in characteristic scales of the microprotrusion and all diode is one of the main difficulties of the task.

2.2 Mathematical model

The electric field potential u in the diode is calculated with the Poisson equation:

$$\begin{aligned} &\Delta u(r,z) = -4\pi\rho(r,z) \\ &\varphi\,|_{cathode} = -U, \quad \varphi\,|_{anode} = 0\,' \end{aligned} \tag{1}$$

where ρ is the space charge density of emitted electrons, which was found by the particle-in-cell method (Hockney & Eastwood, 1988; Birdsall & Langdon, 1985). This equation was solved by a set up method up to decision of a stationary solution at the curvilinear boundary-fitted grid (see Fig. 2). In our electrostatic PIC simulation, each computer particle is a "superparticle" which represents some number of real electrons. The charge of these "superparticles" is not constant and it is defined by expression $q_p = j_{em} \Delta S_i \Delta t$, where j_{em} is the FEE current density, ΔS_i is the elementary area of the emission surface, Δt is time step. The particles start at the cathode microprotrusion, as a result of the FEE process. The particles are then followed, one after the other, during successive time steps. Their trajectory is calculated by Newton's laws

$$\begin{aligned} &z = z_0 + v_z^0 \ \Delta t, \qquad r = r_0 + v_r^0 \ \Delta t, \\ &v_z = v_z^0 + \frac{e}{m} E_z(r,z)\Delta t, \qquad v_r = v_r^0 + \frac{e}{m} E_r(r,z)\Delta t, \end{aligned} \tag{2}$$

where z_0, r_0 and z, r are the posiyion coordinates before and after Δt, v_z^0, v_r^0 and v_z, v_r are the velocities before and after Δt, $E_z = -du/dz$ and $E_r = -du/dr$ are the axial and radial electric field, e and m are the electron charge and mass, respectively.

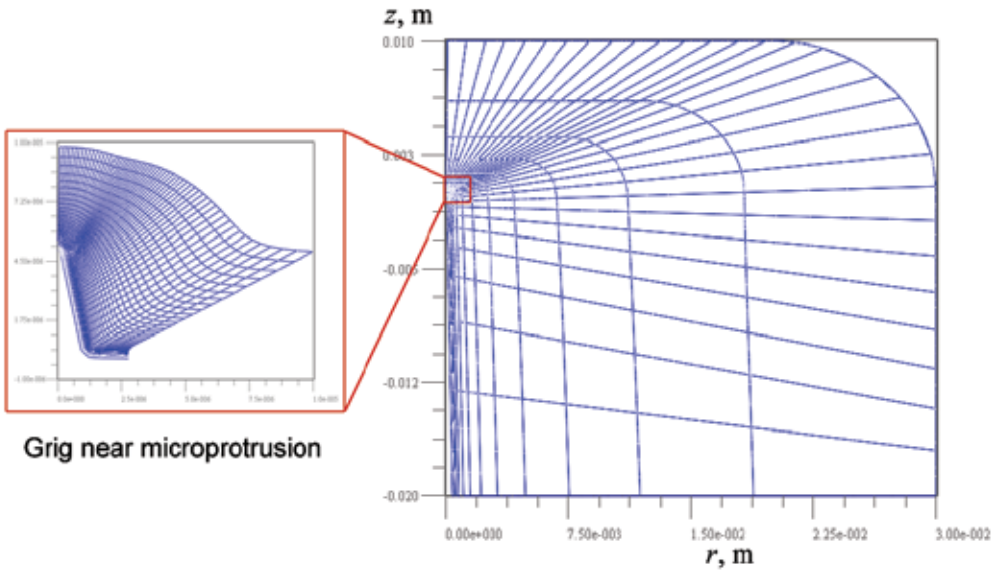


Fig. 2. Discrete representation of the model geometry with a boundary-fitted grid

In a typical electrostatic PIC simulation, for each time step:

1. The charge density $\rho(r,z)$ is obtained by a bilinear weighting of the particles to the spatial curvilinear grid (Seldner & Westermann, 1988).
2. $\rho(r,z)$ is used in Poisson's equation to solve for the electric field $\vec{E} = -\vec{\nabla}u$.
3. E_z and E_r are bilinearly weighted back to each particle position in order to determine the force on each particle.
4. The Newton equations of motion (2) are used to advance the particles to new positions and velocities.
5. The boundaries are checked, and out of bounds particles are removed.

The FEE current density j_{em} was assumed to depend on the self-consistent electric field at the microprotrusion surface in accordance with Miller–Good (MG) approximation (Modinos, 1984):

$$j_{em} = \frac{4\pi me}{h^3}\int_0^{\infty} d\varepsilon \int_0^{\varepsilon} d\varepsilon_n f_{FD}(\varepsilon) D(E_{em}, \varepsilon_n), \tag{3}$$

where $D(E_{em}, \varepsilon_n)$ is the transparency of the potential barrier, E_{em} is the electric field at the cathode, $\varepsilon = (\hbar k)^2/2m$ is the electron energy in the metal, $\varepsilon_n = (\hbar k_n)^2/2m$ is the energy component of an electron in the metal which is "normal to the emission boundary", h ($\hbar = h/2\pi$) is the Plank constant. We assume that $f_{FD}(\varepsilon) = \{1 + \exp((\varepsilon - \varepsilon_F)/k_B T_e)\}^{-1}$ is the

equilibrium Fermi–Dirac function, where ε_F is the Fermi energy, k_B is Boltzmann's constant, T_e = 300°K is the electron temperature.
Within the framework of the MG approximation, the expression for the transmission factor of a barrier has the form (Modinos, 1984):

$$D(E_{\text{em}},\varepsilon_{\text{n}}) = \begin{cases} \left[1+\exp(Q(E_{\text{em}},\varepsilon_{\text{n}}))\right]^{-1}, & \varepsilon_{\text{n}} < \varepsilon_L \\ 1, & \varepsilon_{\text{n}} > \varepsilon_L \end{cases}, \tag{4}$$

$$Q(E_{\text{em}},\varepsilon_{\text{n}}) = \frac{4\sqrt{2}}{3}\left(\frac{m^2e^5}{\hbar^4 E_{\text{em}}}\right)^{1/4} y^{-3/2}\upsilon(y), \tag{5}$$

$$y = \sqrt{e^3 E_{\text{em}}}\Big/\left|\varepsilon_F + \varphi - \varepsilon_{\text{n}}\right| \quad , \tag{6}$$

where $\varepsilon_L = \varepsilon_F + \varphi - 1\Big/\sqrt{2}\ \sqrt{e^3 E_{\text{em}}}$, φ is work function, $\upsilon(y)$ is a function, which is defined through eleptic integrals (Modinos, 1984). The rigorous boundary condition on the cathode surface (3) is another factor that substantially complicates the solution and restricts the choice of the solution technique.

2.3 Results

PIC simulations were performed for a copper cathode in the voltage range U = 5÷500 kV. The results of the numerical calculation of the FEE characteristics are presented in Fig. 3 whose geometric parameters are given in Fig. 1. Figure 3 a) gives the results of computation for the maximal FEE current density on the microprotrusion tip. Figure 3 b) presents the results of the PIC simulations of the total FEE current.
Figure 4 presents the distributions of the electric field strength a) and field emission current density b) over the microprotrusion surface at different voltages. From Fig. 4 it can be seen that at $j_{\text{em}} > 10^7$ A/cm^2 the space charge substantially affects both the magnitude of the field and its distribution over the surface. Note that if the space charge would not been taken into account, the field distribution in Fig. 4 a) would remain constant. Analyzing the curves in Fig. 4 b), it can be noted that the screening effect results in increase of the "effective emission surface".
The results of the numerical calculation of the screening effect of the external electric field by the FEE electron space charge are illustrated in Fig. 5. and Fig. 6. The results obtained show that in the range of high FEE currents ($j_{\text{em}} > 10^9$ A/cm^2) the self-consistent field strength is in fact an order of magnitude lower than its geometric value. Figure 6 presents the respective curves for microprotrusions with different tip radii. From this figure it can be seen that a decrease in tip radius decreases the screening efficiency. This effect is essentially two-dimensional in character. Because the space charge is localized within ~10^{-7} m of the emitting surface, the smaller r_{m}, the lesser is portion of the space charge that participates in the screening of the external field at the microprotrusion tip.
For comparison, the dashed curves in Fig. 3, a) and in Fig. 5, a represent the results that we have obtained by using a quasi-two-dimensional EPD model (Barbour et al., 1963). Analyzing the curves obtained, it can be noted that this model overestimates influence of the SC of the FEE electrons.

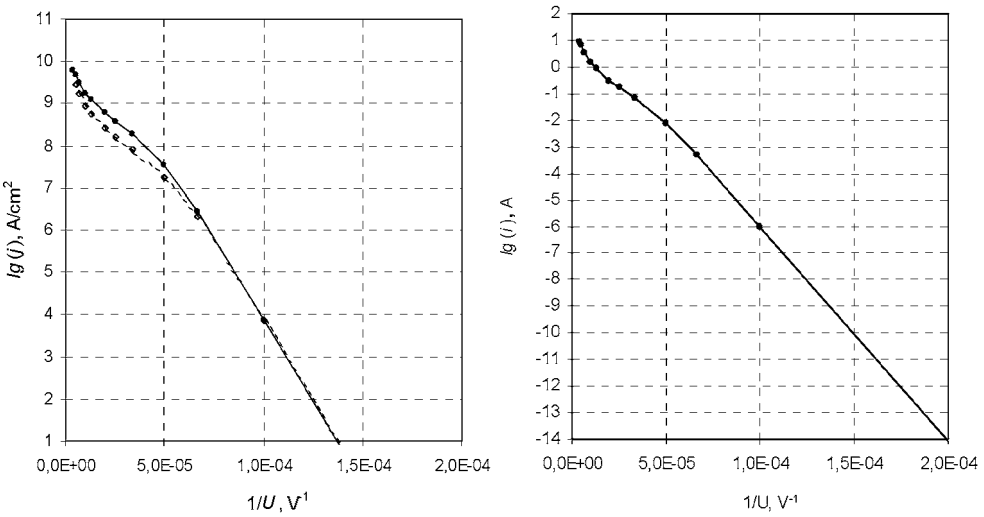


Fig. 3. Calculated current-voltage characteristics for Cu cathode with work function 4.4 eV: a) the FEE current density on the microprotrusion tip (r = 0): 1 – PIC simulations, 2 – numerical calculations within the framework of the EPD model (Barbour et al., 1963); b) PIC simulations of the total FEE current

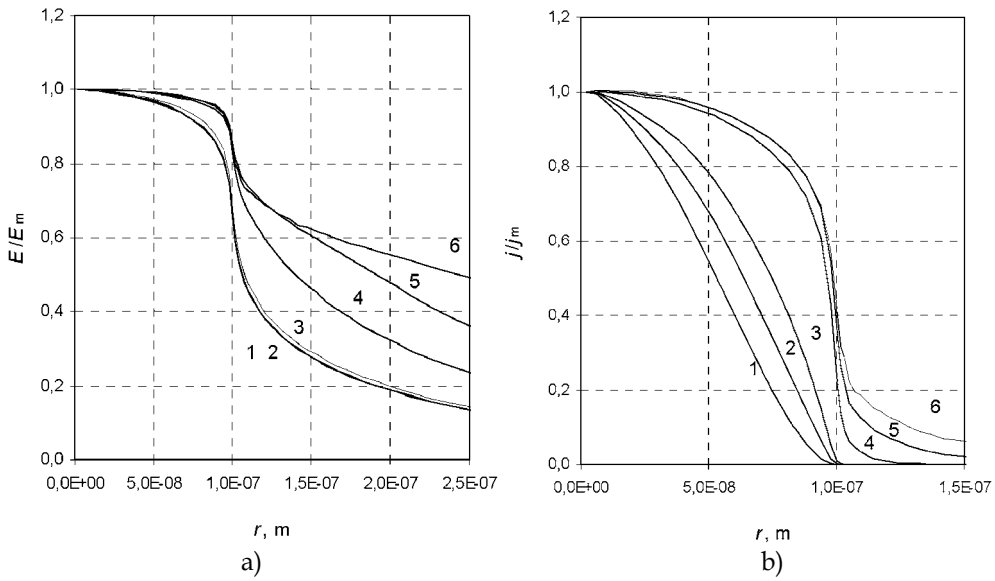


Fig. 4. Electric field strength a) and field emission current density b) distributions over the microprotrusion surface at different voltages (Geometric parameters are given in Fig. 1.): 1 – 5 kV, E_m=1.9 ·10^7 V/cm, j_m=1.2 ·10^{-4} A/cm^2; 2 –15 kV, E_m=5.6 ·10^7 V/cm, j_m=2.7 ·10^6 A/cm^2; 3 – 20 kV, E_m=7.1 ·10^7 V/cm, j_m=3.4 ·10^7 A/cm^2; 4 – 50 kV, E_m=10^8 V/cm, j_m=6 ·10^8 A/cm^2; 5 – 100 kV, E_m=1.2 ·10^8 V/cm, j_m=1.8 ·10^9 A/cm^2; 6 – 250 kV, E_m=1.5 ·10^8 V/cm, j_m=6.2 ·10^9 A/cm^2

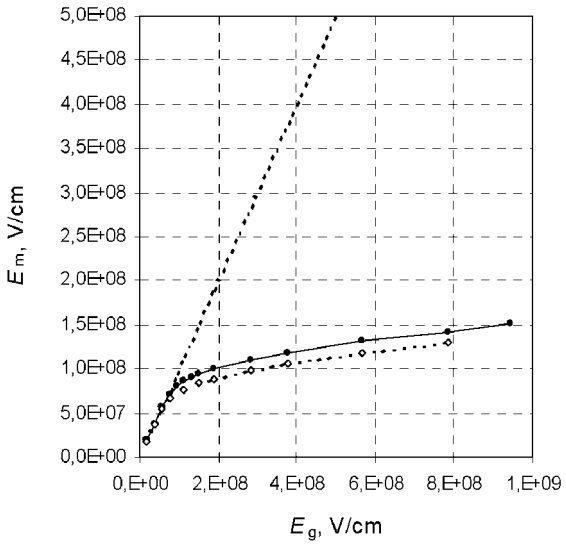


Fig. 5. SC-limited self-consistent electric field at the microprotrusion tip versus geometric field (without taking into account the space charge effect): 1 – geometric field, 2 – PIC simulations, 3 – EPD model (Barbour et al., 1963). Calculated parameters: φ = 4.4 eV, r_c =50 μm, h_m =5 μm, r_m =0.1 μm, Θ =10˚

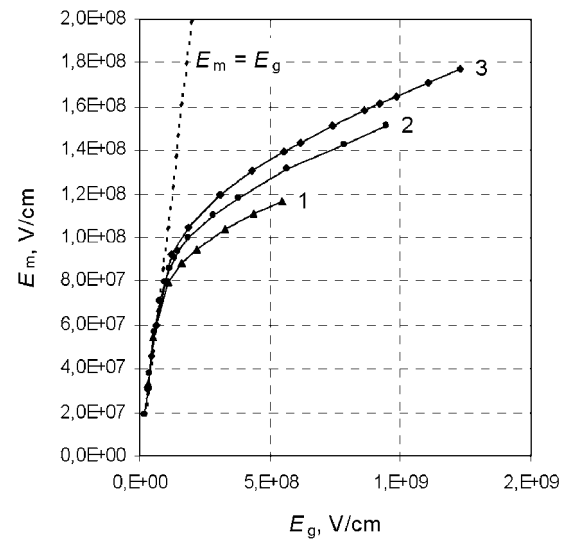


Fig. 6. SC-limited self-consistent electric field at the microprotrusion tip versus geometric field for microprotrusions with different tip radii:
1 – r_m =0.5 μm, 2 – r_m =0.1 μm, 3 – r_m =0.05 μm. Calculated parameters: φ = 4.4 eV, r_c =50 μm, h_m =5 μm, Θ =10˚

2.4 The dimensional effect of the space charge of the emitted electrons on the strength of the self-consistent electric field at the cathode surface

The investigations were performed with the use of above model. The model covers both the trajectory part of the problem and the axially symmetric self-consistent solution of the Poisson equation for the electric field potential over the entire vacuum gap taking into account the SC of emitted electrons. The electron trajectories and the space charge of the emission beam are calculated by the particle-in-cell method using a scheme including macroparticles of varied charge and algorithms of particle coarsening. Coaxial electrode configurations with a point field emitter and the case of emission from a cathode protrusion (see Fig. 1), which are inherent in FEE and EEE investigations, were considered. In both cases, the axial cathode–anode separation was 1 cm. The Dyke model (Dyke et al., 1953) was used for an approximate description of the shape of the point field emitter (see Fig. 7). According to this model, the shape of an emitter prepared by electrolytic etching can be presented rather precisely by the equipotential surface of an electric field produced by a charged orthogonal cone with a sphere on its vertex. The equipotential surface and, hence, the shape of the emitter are specified by three parameters: the radius of the emitter tip, r_0, the radius of the kernel sphere, a, and the order of the Legendre polynomial, n. If the anode is also shaped as an equipotential surface, the solution of the Laplace equation for a system of this configuration is well known (Dyke et al., 1953). Though the field is calculated numerically in our model, the use of this approach allows one to control the procedure of construction of an essentially nonuniform curvilinear computational grid by comparing the accuracy of the numerical solution of the Laplace equation with that of the analytic solution.

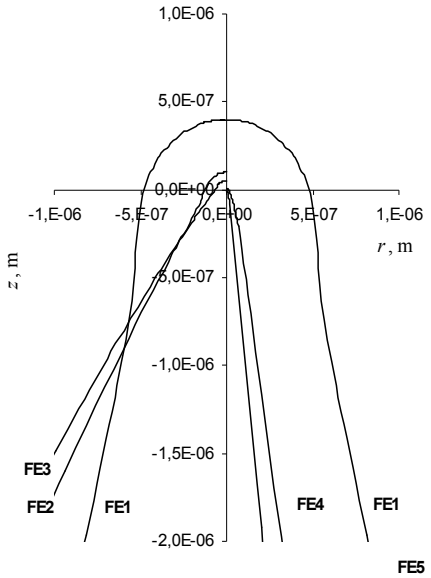


Fig. 7. The shape of model field emitters

The resulting grid is also used for solving the Poisson equation. The parameters of the field emitters that were used in the calculations are given in Table 1. The last column presents the β_0 factor of the emitter, which is calculated numerically from the relation $\beta_0 = E_0 / U$, where E_0 is the field at the tip and U is the cathode–anode potential difference.

N	$r_0 \cdot 10^{-5}$, см	$a \cdot 10^{-5}$, см	n	β_0, см$^{-1}$
FE1	4.0	1.235	0.1	4613
FE2	1.0	0.524	0.2416	4618
FE3	0.5	0.25	0.2835	4599
FE4	0.1	0.02	0.1	$8.0 \cdot 10^4$
FE5	0.01	0.002	0.1	$5.7 \cdot 10^5$

Table 1. Parameters of field emitters

The field strength on the cathode tip calculated as a function of its "geometric" value E_g (Laplace field not taking into account the SC of emitted electrons) and the CVC of the vacuum gap calculated in terms of F–N coordinates for a set of emitters (FE1 through FE3) are shown in Fig. 8 and in Fig. 9, respectively. The emitters of this set are characterized by the same β_0. The data of the respective calculations for emitters FE4 through FE6 having the same cone angle are presented in Fig. 10 and in Fig. 11.

The results obtained suggest that the efficiency of the field screening by the SC of the emission beam depends, in the main, on the emitter radius (linear dimension of the emission area). The smaller the emitter radius, the lower the degree of weakening of the external field at the cathode by the SC of emitted electrons. It should be stressed that this refers both to point emitters and to cathodes with a protrusion. This dimensional effect shows up in the CVC as an increase in current density j_c at which the deviation from the linear F–N characteristic is observed. As can be seen from Fig. 9 and Fig. 11, as the emitter radius is decreased from $4 \cdot 10^{-5}$ to 10^{-7} cm, j_c increases approximately by two orders of magnitude, reaching ~10^9 A/cm^2. The results obtained agree with experimental data for FEE from nanometer protrusions (Pavlov et al., 1975; Fursey et al., 1998).

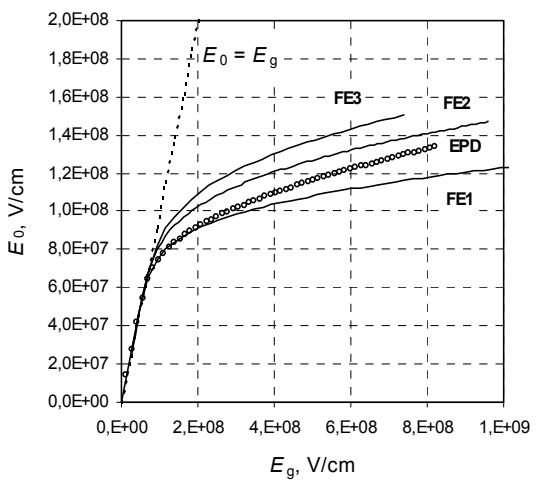


Fig. 8. The field strength at the emitter tip as a function of its geometric value for a set of emitters with the same β_0

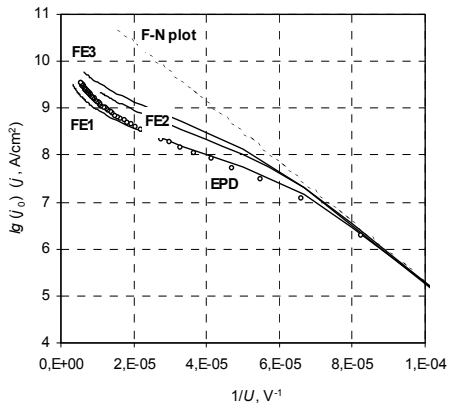


Fig. 9. The FEE current density as a function of the applied voltage for a set of emitters with the same β_0

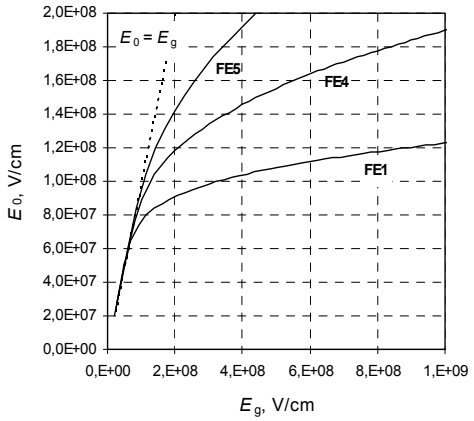


Fig. 10. The field strength at the emitter tip as a function of its geometric value for a set of emitters with the same cone angle

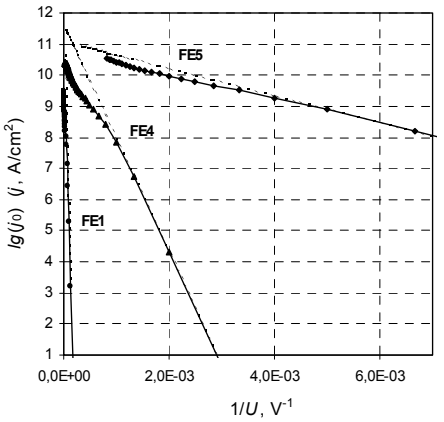


Fig. 11. The FEE current density as a function of the applied voltage for a set of emitters with the same cone angle

In Fig. 8 and Fig. 9, the round markers depict the characteristics under consideration calculated with the use of the EPD model for an emitter with $\beta_0 = 4.6 \cdot 10^3$ cm^{-1}. As can be seen from these figures, good agreement with our results is observed for emitter FE1 with radius $r_0 = 4 \cdot 10^{-5}$ cm up to $j \sim 5 \cdot 10^8$ A/cm^2. Note that the experimental data used in Ref. (Barbour et al., 1963) for comparison were limited to current densities even an order of magnitude lower. However, the results for more intense emission and, especially, for emitters with a smaller radius substantially disagree. Moreover, as emitters FE1 through FE3 have the same parameter β_0, the use of the EPD model yields one solution for this set, notwithstanding that the emitter radius varies within an order of magnitude. Our results evidently show a dependence on emitter radius.

3. Numerical simulation of vacuum prebreakdown phenomena at subnanosecond pulse durations

Considerable advances have recently been achieved in the development of high-current pulsed devices operating on the subnanosecond scale (Mesyats & Yalandin, 2005). In devices of this type, the electron beam is generally produced with the use of an explosive-emission cathode. It should be noted that with the effective duration of the explosive emission process equal to several hundreds of picoseconds, the explosion delay time should be at least an order of magnitude shorter, namely, some tens or even a few picoseconds. It is well known that under the conditions of high vacuum and clean electrodes, explosive electron emission is initiated by the current of field electron emission (Mesyats, 2000). The question of the FEE properties of metals in strong electric fields still remains open from both the theoretical and the experimental viewpoints (Mesyats & Uimanov, 2006). According to the criterion for pulsed breakdown to occur (Mesyats, 2000), to attain picosecond explosion delay times calls for FEE current densities more than 10^9–10^{10} A/cm^2. Investigations performed on the nanosecond scale have shown that at high FEE current densities the electric field strength at the cathode surface is strongly affected by the screening of the electric field with the space charge of emitted electrons (Mesyats, 2000). This is indicated by the deviation of the experimental current-voltage characteristic from the Fowler–Nordheim plot (straight line) in the range of high currents. It was even supposed (Batrakov et al., 1999) that the screening effect may have fatal consequences, so that essentially high current densities which are required to shorten the explosion delay time to picoseconds or even subnanoseconds could not been achieved.

The aim of this section of this work was to investigate the fast processes of heat release and heat transfer that occur in point-shaped microprotrusions of the vacuum-diode cathode within the rise time of a subnanosecond high-voltage pulse. To attain this goal, self-consistent calculations of the field emission characteristics (the current density and the Nottingham energy flux density) were performed taking into account the space charge of emitted electrons (see sec. 2). Since the characteristic times of the processes considered were close to the time of relaxation of the lattice temperature, a two-temperature formulation was used for the model. The current density distribution in a cathode microprotrusion was calculated in view of a finite time of penetration of the electromagnetic field into the conductor.

3.1 Description of the model

Pre-explosive processes occurring on the nanosecond and, the more so, on the microsecond scale, were investigated experimentally and theoretically for rectangular voltage pulses. At

present, this approach can hardly be realized experimentally on the subnanosecond scale. In the experiments described in the available literature (Mesyats & Yalandin, 2005), the pulse shape was near-triangular rather than rectangular with the voltage on the linear section of the leading edge rising at ~10^{15} V/s. Therefore, in this work the voltage at the electrodes was set as a linear function of time, $V(t)$, with dV/dt =1.3·10^{15} V/s. The geometry used in the simulation represented a coaxial diode with 1-cm cathode–anode separation. The cathode was a needle with the tip radius r_c equal to several tens of micrometers. On the cathode surface there was a microprotrusion of height h_m (a few micrometers), tip radius r_m, and cone angle Θ (see Fig. 1.). This cathode geometry takes into account the two-factor field enhancement at the microprotrusion surface which is typical of the electrode systems that were used in the experimental studies of EEE performed by now on the subnanosecond scale.

A two-dimensional two-temperature model which describes the processes of heat release due to surface and bulk sources, the energy exchange between the electron subsystem and phonons, and the heat transfer by electrons has been developed to investigate the prebreakdown phenomena in a cathode microprotrusion for the voltage pulse durations lying in the subnanosecond range. The thermal conductivity of the lattice, with the characteristic times of the problem <10^{-11} s, is neglected.

The electric field potential u in the diode is calculated with the Poisson equation (1). The FEE current density j_{em} was assumed to depend on the self-consistent electric field at the microprotrusion surface in accordance with Miller–Good approximation (3)-(6). The electron T_e and phonon T_p temperature fields in the cathode are calculated with the heat conduction equations:

$$C_V^e \frac{\partial T_e}{\partial t} = \nabla(\lambda_e \nabla T_e) + \frac{j^2}{\sigma} - \mu_T \vec{j} \nabla T_e - \alpha(T_e - T_p) \quad , \tag{7}$$

$$C_V^p \frac{\partial T_p}{\partial t} = \alpha(T_e - T_p) \quad , \tag{8}$$

where C_V^e and C_V^p are the specific heat of the electrons and phonons, respectively, λ_e is the electron thermal conductivity, μ_T is the Thomson factor, σ is the electric conductivity, j is the current density in the cathode, α is the energy exchange factor between the electron subsystem and phonons.

The boundary condition for equation (7, 8) is given by the resulting heat flux at the cathode surface:

$$-\lambda_e \nabla T_e\big|_S = q_N \,, \quad \lambda_p \nabla T_p\big|_S = 0 \,, \tag{9}$$

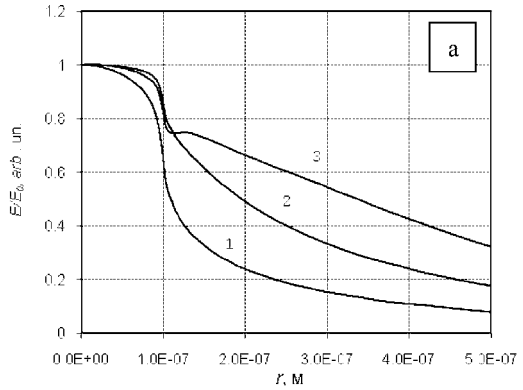
$$q_N = \frac{4\pi me}{h^3} \int_0^\infty \varepsilon \quad d\varepsilon \int_0^\varepsilon d\varepsilon_n f_{FD}(\varepsilon) D(E_{em}, \varepsilon_n) - \frac{j_{em}}{e} \varepsilon_F \,, \tag{10}$$

where q_N is the surface heat release due to Nottingham effect. The other boundary conditions are

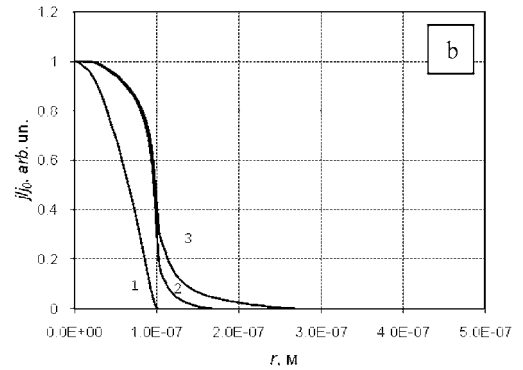
$$\lambda_e \left.\frac{\partial T_e}{\partial r}\right|_S = 0\,, \quad \text{for } r = 0, \quad r \to \infty\,, \tag{11}$$

$$T_e = T_p = T_0\,, \text{for } z \to -\infty\,,\ T_e(r,z)\big|_{t=0} = T_p(r,z)\big|_{t=0} = T_0\,, \tag{12}$$

where T_0 =300 K is the initial homogeneous temperature field for $t = 0$.



a)



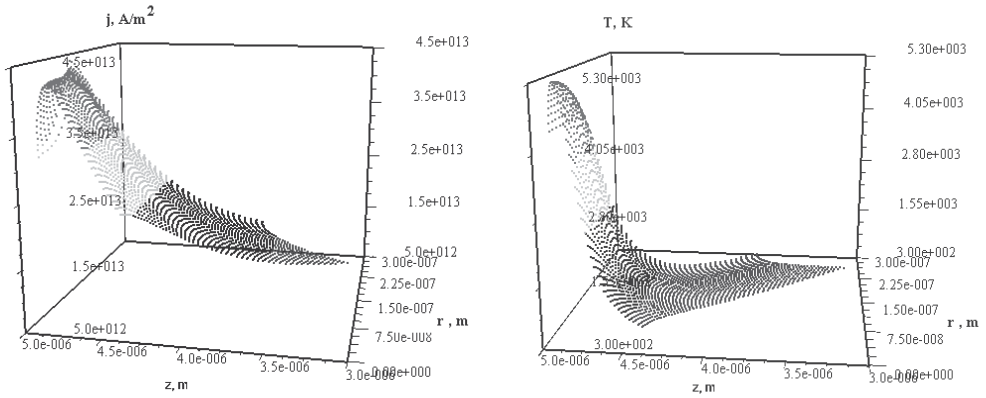
b)

Fig. 12. Distributions of the electric field strength (a) and field emission current density (b) over the microprotrusion surface at different points in time: 1 – t = 2·10^{-16} s, U = 20 kV, T_0 = 300 K, E_0 = 7.3·10^7 V/cm, j_0 = 4.5·10^7 A/cm^2;
2 – t = 3.8·10^{-11} s, U = 70.4 kV, T_0 = 1300 K, E_0 = 1.1·10^8 V/cm, j_0 = 1.0·10^9 A/cm^2;
3 – t = 1.0·10^{-10} c, U = 158 kV, T_0 = 5300 K, E_0 = 1.19·10^8 V/cm, j_0 = 3.2·10^9 A/cm^2

The current density in the cathode $\vec{j} = en_e v_e = (c/4\pi)\, rot\vec{B}$ is determined through the magnetic induction equation:

$$\frac{\partial \vec{B}}{\partial t} = rot\left[\vec{v}_e \vec{B}\right] + \frac{c^2}{4\pi\sigma}\Delta\vec{B}\,. \tag{13}$$

Here, $\vec{B}$ is the magnetic field, e is the electron charge, n_e is the electron density, v_e is the hydrodynamical velocity of the electrons, c is the light velocity.



a) b)

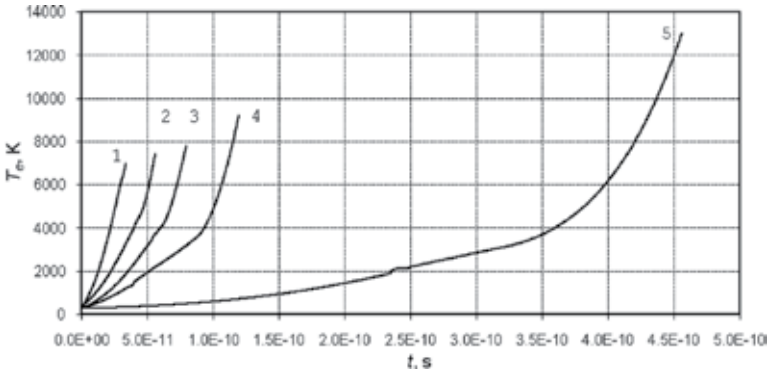
Fig. 13. Distributions of the current density (a) and electron temperature (b) in the Cu microprotrusion for $t = 1.0 \cdot 10^{-10}$ s, $U = 158$ kV, $T_0 = 5300$ K, $E_0 = 1.19 \cdot 10^8$ V/cm, $j_0 = 3.2 \cdot 10^9$ A/cm^2. Geometric parameters are given in Fig. 1

3.2 Results of numerical simulation

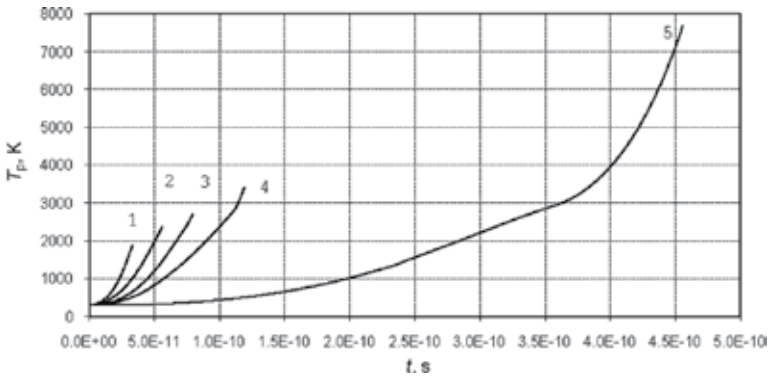
Figure 12 presents the distributions of the electric field strength and field emission current density over the microprotrusion surface at different points in time with a linearly increasing voltage at the electrodes for a copper cathode whose geometric parameters are given in Fig. 1. From Fig. 12 it can be seen that at $j_0 > 10^9$ A/cm^2 the space charge substantially affects both the magnitude of the field and its distribution over the surface. Note that if the space charge would not been taken into account, the field distribution in Fig. 12 a would remain constant. Thus, the screening of the external field by the space charge of emitted electrons substantially levels off the electric field strength at the microprotrusion tip and, accordingly, increases the "effective emission area" (see Fig. 12 b).

With this current density distribution over the microprotrusion surface, the current density is enhanced, as illustrated in Fig. 13 a. Figure 13 b presents the temperature distribution of electrons at the microprotrusion for the same point in time.

The results of a simulation of the microprotrusion heating for different tip radii are given in Fig. 14. The time of relaxation of the lattice temperature $\tau_T^p = C_V^p / \alpha$ and electron temperature $\tau_T^e = C_V^e / \alpha$ are 48 ps and 0.3 ps, respectively. From Fig. 14 it can be seen that thermal instability of the microprotrusion within a time less than $(1–2) \cdot 10^{-10}$ s can develop only for $r_m < 0.1$ μm. In this case, the difference in temperature between the electrons and the lattice can reach 0.5–1 eV. It should be noted that with the geometric parameters of the copper microprotrusion ($r_m < 1$ μm) and the parameters of the pulse used in this simulation ($dV/dt = 1.3 \cdot 10^{15}$ V/s) the effect of a finite time of penetration of the magnetic field in the conductor (skin effect), which is responsible for the nonuniform current density distribution in the microprotrusion and, hence, for its nonuniform heating, is inappreciable. In order that this effect would qualitatively change the spatial distribution of bulk heat sources, shorter times of the processes involved are necessary.



a)



b)

Fig. 14. Time dependences of the maximum electron temperature (a) and lattice temperature (b) in a microprotrusion for a copper cathode of different geometry with r_c = 50 μm, h_m = 5 μm, Θ = 10 deg, and r_m = 0.01 μm (1), 0.03 μm (2), 0.05 μm (3), 0.1 μm (4), and 0.5 μm (5). $dU/dt = 1.3 \cdot 10^{15}$ V/s

4. Initiation of an explosive center beneath the plasma of a vacuum arc cathode spot

Notwithstanding the fact that both the spark and the arc stage of vacuum discharges have been in wide practical use for many years, interest in developing theoretical ideas of the physical phenomena responsible for the operation of this type of discharge is being quickened. In common opinion, the most important and active region in a vacuum discharge is the cathode region. It is our belief that the most consistent and comprehensive model of a cathode spot is the ecton model (Mesyats, 2000). It is based on the recognition of the fundamental role of the microexplosions of cathode regions that give rise to explosive electron emission on a short time scale. The birth of such an explosive center – an ecton – is accompanied by the destruction of a cathode surface region, where a crater is then formed, the appearance of plasma in the electrode gap, and the formation of liquid-metal jets and droplets. An ecton, being an individual cell of a cathode spot, has a comparatively short lifetime (several tens of nanoseconds) (Mesyats, 2000). Therefore, an important issue in this theory is the appearance of new (secondary) explosive centers that would provide for the

self-sustaining of a vacuum discharge. According to (Mesyats, 2000), the most probable reason for the appearance of a new explosive center immediately in the zone of operation or in the vicinity of the previous one is the interaction of a dense plasma with the microprotrusions present on the cathode surface or with the liquid-metal jets ejected from the crater. These surface microprotrusions can be characterized by a parameter β_j which is equal to the ratio of the microprotrusion surface area to its base area and defines the current density enhancement factor. An investigation (Mesyats, 2000) of the development of the explosion of such microprotrusions in terms of the effect of enhancement of the current density of the ions moving from the plasma to the cathode and in view of the Joule mechanism of energy absorption has resulted in the conclusion that for an explosion to occur within 10^{-9} s, it is necessary to have microprotrusions with $\beta_j \geq 10^2$ at the ion current density ~10^7 A·cm^{-2}. This work is an extension of the mentioned model and describes the formation of secondary ectons upon the interaction of a dense plasma with cathode surface microprotrusions.

In the general case, the charge particle flow that closes onto a microprotrusion consists of three components: an ion flow and an electron flow from the plasma and a flow of emission electrons (Ecker, 1980; Hantzsche, 1995; Beilis, 1995). Each of these flows carries both an electric charge and an energy flux, forming a space charge zone at short distances from the cathode surface and giving rise to an electric field E_c at the cathode. In (He & Haug, 1997) the initiation of a cathode spot was investigated for the ion current j_i and the electric field at the cathode E_c specified arbitrarily from a "black box" and with an artificially created spatially homogeneous "plasma focus" of radius 10 μm on a plane cathode. It has been shown that the cathode heating by incident ions and the enhancement of the electric field E_c by the ion space charge reduce the critical field at which the process of thermal run-away and overheating below the surface starts developing. It should however be noted that the least times of cathode spot initiation obtained in (He & Haug, 1997) are longer than 1 μs. On the other hand, according to the ecton model of a cathode spot (Mesyats, 2000) and to the experimental data (see, for example (Juttner, 2001)), the cathode spot phenomena have an essentially nonstationary and cyclic character with the characteristic time scale ranging between 10^{-9} and 10^{-8} s.

Thus the goal of this section is to investigate the formation of secondary explosive centers upon the interaction of the plasma of a vacuum arc cathode spot with cathode surface microprotrusions (Uimanov, 2003).

4.1 Description of the model of the Initiation of an explosive center

The problem statement and task geometry

Figure 15 presents the model geometry of the problem. The shape of the microprotrusion surface is specified by the Gauss function $z_S = h\exp(-(r/d)^2)$, where h is the height of the microprotrusion, d specifies the base radius r_m that is determined for $z = 0.1h$. We shall further characterize the geometry of a microprotrusion by a current density enhancement factor $\beta_j = S/\pi r_m^2$, where S is the surface area of the microprotrusion. In terms of this model, we assume that over the cathode surface there is a cathode spot plasma with an ion density n_i and an electron temperature T_e at the sheath edge. The quantities n_i and T_e are the problem parameters and, according to (Shmelev & Litvinov, 1998), they depend on the distance from the active explosive center.

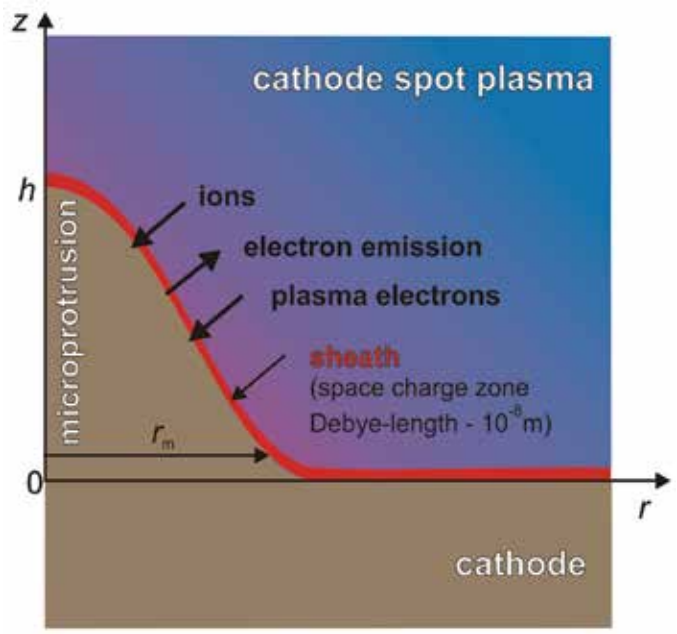


Fig. 15. Task geometry and a schematic description of the 2D model in a cylindrical symmetry

In view of the fact that the width of the space charge layer is much smaller than the characteristic dimensions of the microprotrusion, the layer parameters are considered in a one-dimensional (local) approximation.

The temperature field

The temperature field in the cathode is calculated with the heat conduction equation:

$$c_p \rho \frac{\partial T}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left(r\lambda \frac{\partial T}{\partial r} \right) + \frac{\partial}{\partial z}\left(\lambda \frac{\partial T}{\partial z} \right) + \frac{j^2}{\sigma}, \tag{14}$$

where c_p is the specific heat at constant pressure, λ is the thermal conductivity, ρ is the mass density, σ is the electric conductivity, j is the current density in the cathode. The parameters c_p, λ and σ are considered as function of temperature $T(r,z,t)$ (Zinoviev, 1989). The boundary condition for equation (14) is given by the resulting heat flux at the cathode surface:

$$-\lambda \nabla T\big|_S = q_S, \tag{15}$$

where $q_S = q_N + q_i$ is the sum of the Nottingham effect q_N (see eq. 10) and ion impact heating q_i (evaporation cooling does not noticeably affect the final results). The other boundary conditions are

$$\lambda \frac{\partial T}{\partial r} = 0 \text{, for } r = 0, \quad r \to \infty \text{, } T = T_0 \text{, for } z \to -\infty, \tag{16}$$

where T_0 =300 K is the initial homogeneous temperature field for $t = 0$.

The Joule heating

The Ohm's electric potential U and current density $\vec{j} = -\sigma \vec{\nabla} U$ in the cathode is determined through the continuity equation:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\sigma \frac{\partial U}{\partial r} \right) + \frac{\partial}{\partial z}\left(\sigma \frac{\partial U}{\partial z} \right) = 0 , \tag{17}$$

with boundary condition at the cathode surface:

$$-\nabla U\big|_S = j_s / \sigma . \tag{18}$$

Here $j_s = j_{\text{em}} + j_i$ is the total current density at the cathode surface, j_{em} is the electron emission current density and j_i is the current density of the ions moving from the plasma to the cathode. The other boundary conditions are

$$\sigma \frac{\partial U}{\partial r} = 0 \quad \text{, for } r = 0, \quad r \to \infty , \qquad U = 0 \text{, for } z \to -\infty , \tag{19}$$

The plasma-surface interaction

To calculate j_i, it is assumed that the ions are treated as monoenergetic particles, entering the sheath edge with Bohm's velocity and all ions recombine at the cathode surface. Then the expression for j_i can be written in the form:

$$j_i = Zen_i \sqrt{\frac{kT_e}{m_i}} , \tag{20}$$

where Z is the mean ion charge, m_i is the ion mass. The power density input into the cathode surface from ion impact heating (Mesyats & Uimanov, 2006) is $q_i = j_i \bar{U}$, where $Ze\bar{U} = ZeV_{\text{c}} + eV_i - Z\varepsilon_{\text{em}}$. Here V_{c} is the cathode fall potential, V_i is the averaged ionization potential, ε_{em} is the averaged energy per emitted electron. The electron emission characteristics j_{em} and $\varepsilon_{\text{em}} = q_N / (j_{em} / e)$ are calculated numerically in the MG approximation (see sec. 2.2 and sec. 3.1). Because of the high temperatures, the temperature drift of the chemical potential of the electron system inside the cathode is taken into account (see, for example (Klein et al., 1994)). To calculate the electric field at the cathode, the Mackeown-like equation is used, taking into account the electron flow from the spot plasma to the cathode (Mackeown, 1929; Mesyats & Uimanov, 2006; Beilis, 1995):

$$E_{\text{em}}^2 = \frac{4}{\varepsilon_0}\left[j_i(n_i, T_e)\sqrt{\frac{m_i V_{\text{c}}}{2Ze}} \left\{ \sqrt{1 + \frac{kT_e}{2ZeV_c}} - \sqrt{\frac{kT_e}{2ZeV_c}} - \sqrt{\frac{kT_e}{2ZeV_c}}\left(1 - \exp\left(-\frac{eV_{\text{c}}}{kT_e} \right) \right) \right\} - \right.$$
$$\left. - j_{\text{em}}(E_{\text{c}}, T_S)\sqrt{\frac{m_e V_{\text{c}}}{2e}} \right] , \tag{21}$$

where T_S is the cathode surface temperature.

In conclusion of this section, we explain in more detail how the model proposed takes into account the contribution of the electron flow from the plasma to the cathode. If we assume

that the velocity distribution of the plasma electrons at the sheath edge is a Maxwellian one, we arrive at the statement that only the electrons whose velocities are higher than $\sqrt{2eV_c / m_e}$ make a contribution to the current of the electrons 'counterdiffusing' from the quasineutral plasma to the cathode, j_{ep}. According to (Hantzsche, 1995), we have $j_{ep} = -Zen_i\sqrt{kT_e / 2\pi m_e}\exp(-eV_c / kT_e)$. Then, in view of (20), the ratio of this contribution to the ion current density takes the form $|j_{ep}| / j_i = \sqrt{m_i / 2\pi m_e}\exp(-eV_c / kT_e)$. For the parameters used in this work ($\sqrt{m_i / m_e} \approx 340$, $eV_c / kT_e = 8$), we have $|j_{ep}| / j_i = 4.6\cdot10^{-2}$. Therefore, in the boundary condition Eq. (18), we may neglect the contribution j_{ep} to the total current at the surface of the microprotrusion. The small ratio of the electron current to the current of the ions arriving at the cathode from the spot plasma permits us to ignore as well the energy flux density of these electrons, q_{ep}, in the general balance of the surface heat sources in the boundary condition Eq. (15). With the parameters used, we have $q_{ep} / q_i \approx (0.1\div2)\times10^{-2}$. Thus, in the case under consideration, the contributions of j_{ep} and q_{ep} to the current and energy balance at the cathode surface can be neglected. At the same time, it should be stressed that the effect of the electron flow from the plasma to the cathode is essential in calculating the characteristics of the space charge sheath (see Eq. (21)). If we take account of the effect of the space charge of this flow, we obtain that E_c noticeably decreases. This results in a substantial change in the rate of the development of thermal instability because of the strong dependence of the emission characteristics on E_c.

Microprotrusion parameters					Explosion delay time t_d, ns	
N	h, μm	d, μm	r_m, μm	β_j	$j_i = 5.6\cdot10^{10}$ A/m²	$j_i = 1.1\cdot10^{11}$ A/m²
1	0.5	0.312	0.5	1.4	-	10
2	1	0.312	0.5	2.2	-	1.5
3	1.5	0.312	0.5	3	16.3	0.8
4	1.75	0.312	0.5	3.52	5.2	0.71
5	2	0.312	0.5	4	3.55	0.68
6	3	0.312	0.5	5.9	2.26	0.57
7	5	0.312	0.5	9.74	1.5	0.33

Table 2. A set of geometrical parameters of the microprotrusions and the obtained explosion delay time

4.2 Simulation of the microprotrusion heating

Computations were performed for a copper cathode with the following arc parameters: V_c=16 V, eV_i=18 eV, Z=2. For the initial conditions of the problem, the characteristics of the plasma (n_i, T_e) at the sheath edge and the microprotrusion geometry (h, d) were specified. The computations were performed for a set of geometrical parameters of the microprotrusions, which are submitted in the Table 2.

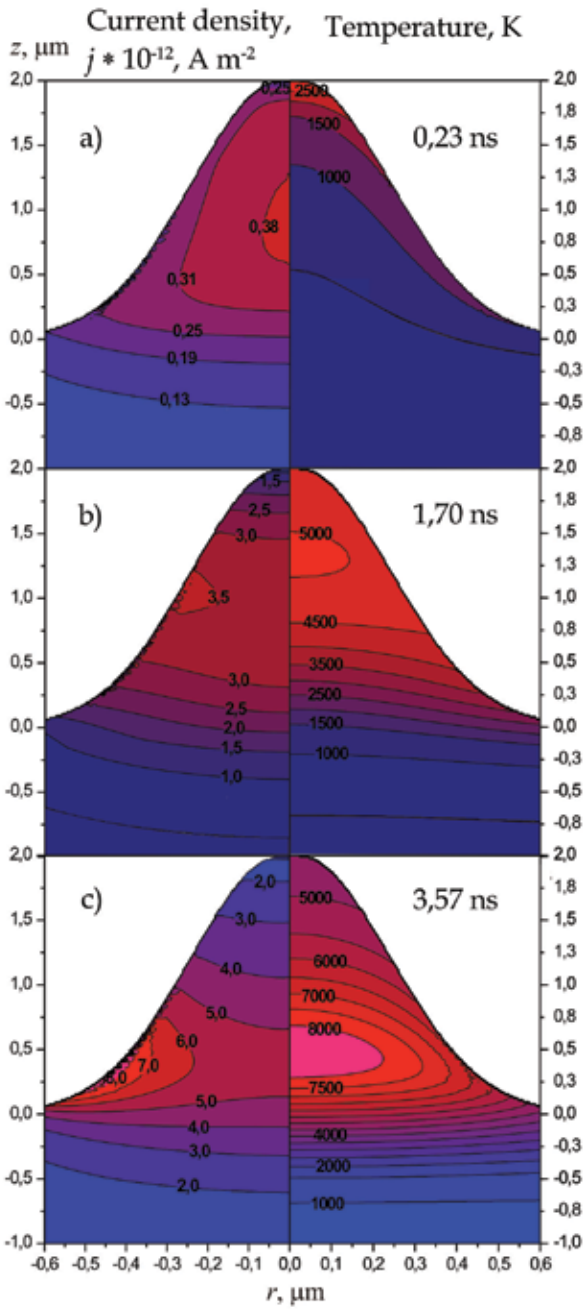


Fig. 16. Space distribution of temperature and current density modulus: a) t = 0.23 ns, b) t = 1.9 ns, c) t = 6.2 ns

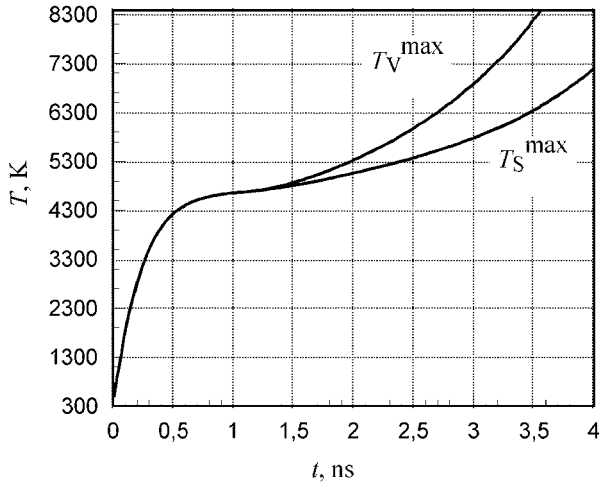


Fig. 17. The temperature evolution of the microprotrusion with β_j =4. Cathode spot plasma parameters: n_i =10^{26} m^{-3} T_e=2 eV (j_i=5.6·10^{10} A/m^2). T_S^{max} is the surface maximum temperature and T_V^{max} is the bulk maximum temperature that is reached below the surface due to emission cooling

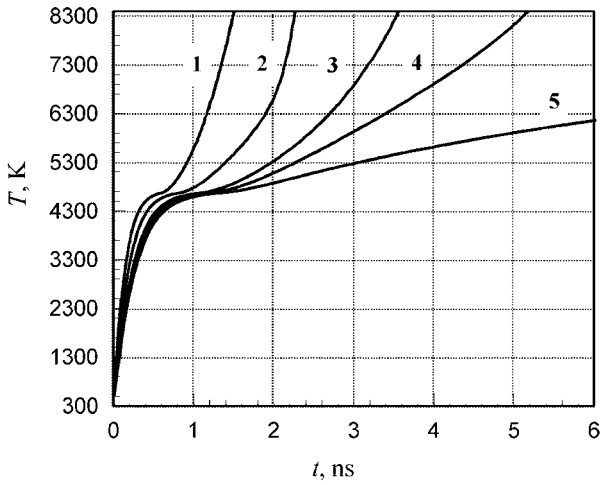


Fig. 18. The influence of the microprotrusion geometry on the initiation of the explosive stage: 1 - β_j = 9.74; 2 - β_j = 5.9; 3 - β_j = 4; 4 - β_j = 3.52; 5 - β_j = 3. Cathode spot plasma parameters: n_i =10^{26} m^{-3} T_e=2 eV (j_i=5.6·10^{10} A/m^2)

The computation is performed until the maximum temperature in the protrusion reaches a critical temperature T_{cr}. =8390 K As this happens, the model becomes inoperative, and the process goes to the explosive phase of the development of an ecton.

Figure 16 gives the results of computation for the temperature field and the spatial distribution of the current density modulus.

The simulation has shown that the heating of a microprotrusion can be subdivided by convention into two stages. At the first stage (see Fig. 16 a)), where the cathode is still comparatively cold, the surface heating due to ion impact prevails. In this case, the microprotrusion behaves as if it were a collecting thermal lens. The Joule heating at this stage was inessential . As the temperature reaches ~3500÷4000 K, the emission current density increases substantially and intense surface cooling begins, and this can be associated with the onset of the second stage of heating (see Fig. 16 b)). At this point, the current density maximum is in the bulk of the microprotrusion and, hence, this is the region where intense Joule heat release begins. This region is responsible for the maximum temperature in the cathode at any subsequent point in time (see Fig. 16 c). Then the maximum surface temperature shifting toward the protrusion base. The current density maximum also tends to move to this region since the current "makes attempts" to bypass the high-temperature region. After a time, this surface region has the highest surface temperature and emission current density. However, the most intense heat release occurs, as earlier, in the microprotrusion bulk, and thus a highly overheated region is formed which is surrounded by a not so hot surface. Figure 17 presents the time dependence of the temperature being a maximum throughout the microprotrusion and for the surface temperature that underlie the character of the above process of the heating of a microprotrusion.

Figure 18 was obtained with the same ion current density, but with different values for β_j (the microprotrusion geometry). It clearly shows the development of the thermal run-away regime and the initiation of the explosive stage.

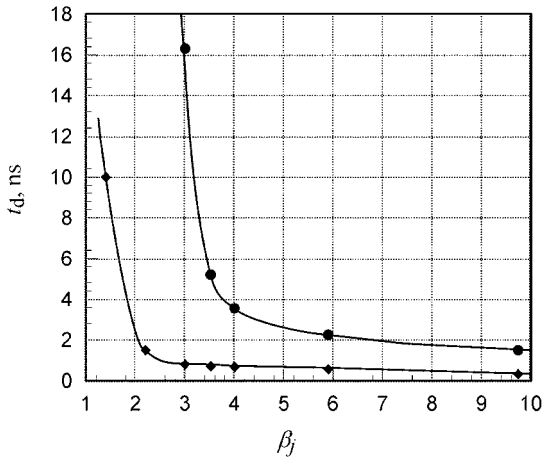


Fig. 19. Explosion delay time *vs* the current density enhancement factor β_j : • - n_i=10^{26} m^{-3} T_e=2 eV (j_i=5.6·10^{10} A/m^2), ♦ - n_i =1.8·10^{26} m^{-3} T_e=2 eV (j_i=1.1·10^{11} A/m^2)

The results of the computations performed are combined in Fig. 19 and Table 2 where the explosion delay time t_d (heating time from 300 K to T_{cr}) is presented as a function of the microprotrusion geometry for a varied plasma density (ion current density).

It should be stressed that there is a range of comparatively small values of β_j where the given mechanism of the formation of secondary ectons can ensure the birth of new cathode spot cells upon the interaction of the plasma generated by active cells with cathode surface microprotrusions.

5. Conclusion

The effect of the space charge of the emitted electrons on the strength of the self-consistent electric field at the surface of a pointed microprotrusion and field emitter has been investigated for the first time in the framework of a two-dimensional axisymmetric statement of the problem. Based on the particles in cells method, a model has been developed and self-consistent calculations of the electric field and of the field emission characteristics of the cooper cathode taking into account the space charge of emitted electrons have been performed for the range of emitter tip radius from ~10^{-4} cm to ~1 nm. For the geometry under investigation it has been shown that the space-charge screening of the external field is substantially less pronounced for emitters whose tip radii are comparable to the size of the region where the space charge is mostly localized. As for emitters having nanometer tip radii, their CVCs remain linear in the F–N coordinates up to FEE current densities of ~10^9 A/cm^2. Despite the significant screening of the external field at high FEE current densities, the emission current density for microprotrusions with tip radii r_m < 0.1 μm can reach ~10^{10} A/cm^2. Based on the criterion for pulsed breakdown (Mesyats, 2000), it can been shown that, in view of Joule heating, this current density suffices for the FEE-to-EEE transition to occur within less than 10^{-10} s.

A two-dimensional, two-temperature model has been developed to describe the prebreakdown phenomena in a cathode microprotrusion at nanosecond durations of the applied voltage pulse. The simulation procedure includes (i) a particle-in-cell simulation to calculate the self-consistent electric field at the cathode and the field-emission characteristics of the cathode; (ii) calculations of the current density distribution in the cathode microprusion in view of a finite time of electromagnetic field penetration in the conductor; (iii) calculations of the electron temperature based on the heat equation taking into account volumetric (Joule–Thomson effect) and surface (Nottingham effect) heat sources, and (iv) calculations of the lattice temperature based on the heat equation taking into account the finite time of electron-phonon collisions. A numerical simulation performed for a copper cathode for voltage pulses with ~10^{15} V/s rise rates has demonstrated that (i) the screening of the external field by the space charge of emitted electrons has the result that the electric field strength levels off approaching to that at the microprotrusion tip, and this gives rise to a region inside the microprotrusion where the current density is about twice the maximum field emission current density at the tip; (ii) the electron temperature can be greater than the lattice temperature by 0.5–1 eV at the onset of the explosive metal-plasma phase transition; (iii) with a 5-μm characteristic height of microprotrusions on a point cathode whose radius of curvature is 50 μm the field emission-to-explosive emission transition can occur within 100÷200 ps only for microprotrusions with a tip radius no more than 0.1 μm.

A two-dimensional nonstationary model of the initiation of new explosive centers beneath the plasma of a vacuum arc cathode spot has been developed. In terms of this model, the plasma density and electron temperature that determine the ion current from the plasma to the microprotrusion and the microprotrusion geometry were treated as the external parameters of the problem. The process of heating of a cathode surface microprotrusion, for which both a surface irregularity resulting from the development of a preceding crater and the edge of an active crater, which may be a liquid-metal jet, can be considered, has been simulated numerically. Based on the computation results, one can make the following conclusions:

i. The heating of a microprotrusion gives rise to a strongly overheated region in the protrusion bulk. Hence, an expansion of such a microregion of the cathode, being in an extreme state, should be explosive in character.
ii. Taking into account the ion impact heating and the electric field of the space charge layer near the cathode surface ensure the "triggering" heat flux power necessary for the development of the Joule heating of the microprotrusion followed by it explosion at reasonable values of the ion current ($j_i < 10^7$ A·cm^{-2}) and of the geometric parameters of the microprotrusion ($\beta_j < 10$).

6. Acknowledgment

The author would like to thank Academician G. A. Mesyats, who provided encouragement and stimulating discussion.
The work was partial supported by the Russian Fundamental Research Foundation under Awards 10-08-00517, 08-02-00720 and the integration project of the Presidium UB RAS No. 09-C-2-10002.

7. References

Barbour, J.P.; Do1an, W.W.; Tro1an, J.K.; Martin, E.E. & Dyke, W.P. (1963). *Phys. Rev.*, Vol. 92, (1963), p. 45

Batrakov, A.V.; Pegel, I.V. & Proskurovsky, D.I. (1999). On the screening of the electric field at the cathode surface by an electron space charge at intense field emission, *IEEE Trans. Dielectrics El.*, Vol. 6, No. 4, August 1999, pp. 436-440, ISSN 1070-9878

Batrakov, A.V.; Pegel, I.V. & Proskurovsky, D.I. (1999). *Rus. Pisma J. Tech. Physics*, Vol. 25, (1999), p. 78

Beilis, I.I. (1995). Theoretical Modeling of Cathode Spot Phenomena, In: *Handbook of Vacuum Arc Science and Technology,* Boxman, R.L. et. al. (Ed.), pp. 208-256, Noyes, New Jersey

Birdsall, C.K. & Langdon, A.B. (1985). *Plasma Physics, via Computer Simulation*, McGraw-Hill Book Company

Dyke, W.P. & Trolan, J.K. (1953). *Phys. Rev.*, Vol. 89, (1953), p. 799

Dyke, W.P.; Tro1an, J.K.; Do1an, W.W. & Barbour, J.P. (1953). *J. Appl. Phys.*, Vol. 24, p. 570

Ecker, G. (1980). Theoretical aspects of the vacuum arc, In: *Theory and Aplication,* Lafferty, J.M. (Ed.), pp.228-320, Wiley, New York

Forbes, R.G. (2001). *Sol. St. Electron*, Vol. 45, (2001), p. 779

Fursey, G.N.; Baskin, L.M.; Glasanov, D.V. et al. (1998). *J. Vac. Sci. Technol. B*, Vol. 16, (1998), p. 232

Fowler, R.H. & Nordheim, L. (1928). *Proc. Roy. Soc.*, Vol. 119, (1928), p. 173

Guillorn, M.A.; Yang, X.; Melechko, A.V. et al. (2004). *J. Vac. Sci. Technol. B*, Vol. 22, (2004), p. 35

Handbook of Vacuum Arc Science and Technology (1995). Boxman, R.L.; Martin, P.J. & Sanders, D.M. (Ed.), Noyes Publications, Park Ridge

Hantzsche, E. (1995). Theories of Cathode Spots In: *Handbook of Vacuum Arc Science and Technology,* Boxman, R.L. et. al. (Ed.), pp.151-208, Noyes, New Jersey

He, Z.-J. & Haug, R. (1997). Cathode spot initiation in different external conditions, *J. Phys. D: Appl. Phys.*, Vol. 30, No. 4, Feb. 1997, pp. 603-613

Hockney, R.W. & Eastwood, J.W. (1988). *Computer Simulation Using Particles*, IOP Publishing, Bristol

Juttner, B. (2001). Cathode spots of electric arcs, *J. Phys. D: Appl. Phys.*, Vol. 34, No. 17, Sep. 2001, pp. R103-R123

Klein, T.; Paulini, J. & Simon, G. (1994). Time-resolved description of cathode spot development in vacuum arcs, *J. Phys. D: Appl. Phys.*, Vol. 27, No. 9, Sep. 1994, pp. 1914-1921

Lyubimov, G.A. & Rakhovskii, B.I. (1978). The cathode spot of a vacuum arc, *Physics-Uspekhi*, Vol. 21, pp. 693-718

Mackeown, S.S. (1929). The Cathode Drop in an Electric Arc, *Phys. Rev.*, Vol. 34, No. 4, Aug. 1929, pp. 611–614

Mesyats, G.A. & Uimanov, I.V. (2006). On the limiting density of field emission current from metals, *IEEE Trans. Diel. Elec. Insul., V*ol. 13, (Feb. 2006), pp. 105–110

Mesyats, G.A. & Uimanov, I.V. (2008). Numerical Simulation of Vacuum Prebreakdown Phenomena in a Cathode Microprotrusion at Subnanosecond Pulse Durations, *Proc. of XVIIIth Int. Symp. On Disch. and Electr. Insul. in Vac.*, pp. 17-20, ISBN 978-973-755-382-9, Bucharest, Romania, Sept. 2008, MATRIX ROM, Bucharest

Mesyats, G.A. & Yalandin, M.I. (2005). High-power picosecond electronics, *Physics-Uspekhi*, Vol. 48, No. 3, Mar. 2005, pp. 211–229

Mesyats, G.A. (2000). *Cathode Phenomena in Vacuum Discharge: the Breakdown, the Spark and the Arc*, Nauka Publisher, ISBN 5-02-022567-3, Moscow

Modinos, A. (1984). *Field, Thermionic and Secondary Electron Emission Spectroscopy*. Plenum Press, New York

Nordheim, L. (1929). *Physikalische Zeitschrift*, Vol. 30, (1929), p. 117

Pavlov, V.G. (2004). *Rus. J. Tech. Physics* , Vol. 74, (2004), p. 72

Pavlov, V.G.; Rabinovich, A.A. & Shrednik, V.N. (1975). *Rus. J. Tech. Physics*, Vol. 45, (1975), p. 2126

Seldner, D. & Westermann, T. (1988). Algorithms for Interpolation and Localization in Irregular 2D Meshes, *J. of Comp. Phys.*, Vol. 79, (Nov. 1988), pp. 1-11

Shkuratov, S.I.; Barengolts, S.A. & Litvinov, E.A. (1995). Heating and Failure of Niobium Tip Cathode due to a High-density Pulsed Field Electron Emission Current, *J. Vac. Sci. Technol. B*, Vol. B13, No. 5, 1995, pp. 1960-1967

Shmelev, D.L. & Litvinov, E.A. (1998). Computer simulation of ecton in vacuum arc, *IEEE Trans. Dielectrics El.*, Vol. 6, No. 4, August 1999, pp. 441-444, ISSN 1070-9878

Shrednik, V.N. (1974). In: *Cold Cathodes*, pp. 165-190, Sov. Radio, Moscow

Spindt, C.A. (1968). *J. Appl. Phys*, Vol. 39, (1968), p. 3504

Stern, T.E.; Gos1ing, B.S. & Fowler R.H. (1929). *Roy. Soc. Proc.*, Vol. A 124, (1929), p. 699

Uimanov, I.V. (2003). A Two-Dimensional Nonstationary Model Of The Initiation Of An Explosive Center Beneath The Plasma Of A Vacuum Arc Cathode Spot, *IEEE Transaction on Plasma Science*, Vol. 31, N. 5, 2003, pp. 822-826

Uimanov, I.V. (2008). PIC Simulation of the Electric Field at a Cathode with a Surface Microprotrusion Under Intense Field Emission, *Proc. of XVIIIth Int. Symp. On Disch. and Electr. Insul. in Vac.*, pp. 29-31, Bucharest, Romania, (2008)

Uimanov, I.V. (2010). The Dimensional Effect of the Space Charge of the Emitted Electrons on the Strength of the Self-consistent Electric Field at the Cathode Surface, *Proc. 16th International Symposium on High Current Electronics*, Tomsk, Russia (to be published) (2010)

Zinoviev, V.E. (1989). *Thermal Properties of Metals at High Temperatures. Reference Book*. Metallurgia, Moscow

3

3-D Quantum Numerical Simulation of Transient Response in Multiple-Gate Nanowire MOSFETs Submitted to Heavy Ion Irradiation

Daniela Munteanu and Jean-Luc Autran
IM2NP Laboratory,
CNRS (Centre National de la Recherche Scientifique), Aix-Marseille University
France

1. Introduction

The bulk MOSFET scaling has recently encountered significant limitations, mainly related to the gate oxide (SiO_2) leakage currents (Gusev et al., 2006; Taur et al., 1997), the large increase of parasitic short channel effects and the dramatic mobility reduction (Fischetti & Laux, 2001) due to highly doped Silicon substrates precisely used to reduce these short channel effects. Technological solutions have been proposed in order to continue to use the "bulk solution" until the 32 nm ITRS node (ITRS, 2009). Most of these solutions envisage the introduction of high-permittivity gate dielectric stacks (to reduce the gate leakage, (Gusev et al., 2006; Houssa, 2004), midgap metal gate (to suppress the Silicon gate polydepletion-induced parasitic capacitances) and strained Silicon channel (to increase carrier mobility (Rim et al., 1998). However, in parallel to these efforts, alternative solutions to replace the conventional bulk MOSFET architecture have been proposed and studied in the recent literature. These options are numerous and can be classified in general according to three main directions: (i) the use of new materials in the continuity of the "bulk solution", allowing increasing MOSFET performances due to their dielectric properties (permittivity), electrostatic immunity (SOI materials), mechanical (strain), or transport (mobility) properties; (ii) the complete change of the device architecture (e.g. Multiple-Gate devices, Silicon nanowires MOSFET) allowing better electrostatic control, and, as a result, intrinsic channels with higher mobilities and currents; (iii) the exploitation of certain new physical phenomena that appear at the nanometer scale, such as quantum transport, substrate orientation or modifications of the material band structure in devices/wires with nanometer dimensions (Haensch et al., 2006; Hiramoto et al., 2006).

Multiple-Gate nanowire MOS transistors (Fig. 1) are now widely recognized as one of the most promising solutions for meeting the roadmap requirements in the deca-nanometer scale (Park & Colinge, 2002). A wide variety of architectures, including planar Double-Gate (DG) (Frank et al., 1992; Harrison et al, 2004), Vertical Double-Gate, Triple-Gate (Tri-gate) (Guarini et al., 2001; Park & Colinge, 2002), FinFET (Choi et al., 2001; Kedzierski et al., 2002), Omega-Gate (Ω -Gate) (Park et al., 2001), Pi-Gate (π -Gate) (Yang et al., 2002), Δ-channel SOI MOSFET (Jiao & Salama, 2001), DELTA transistor (Hisamoto et al., 1989), Gate-All-Around (GAA) (Colinge et al., 1990; Park & Colinge, 2002), Rectangular or Cylindrical nanowires

(Jimenez et al., 2004), has been proposed in the literature. These structures exhibit a superior control of short channel effects resulting from an enhanced electrostatic coupling between the conduction channel and the surrounding gate electrode. It has been shown that the electrostatic control is enhanced when increasing the "Equivalent Number of Gates" (EGN) from 2 (for Double-Gate devices, Fig. 1) to 4 (for Gate-All-Around devices where the gate electrode is wrapped around the entire channel, Fig. 1) (Bescond et al., 2004).

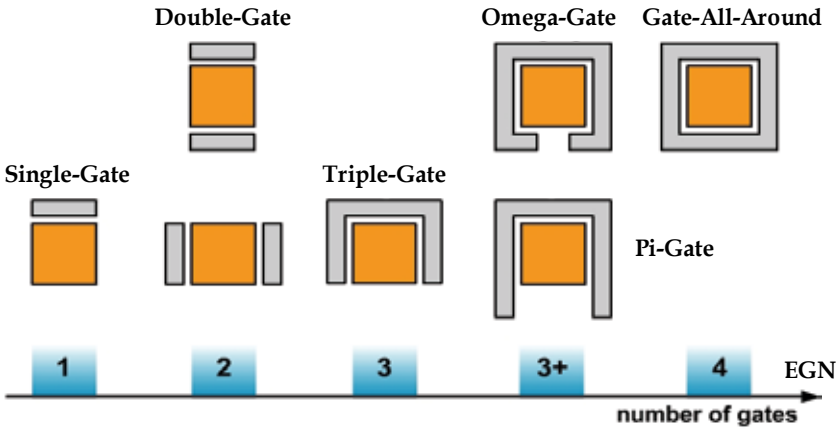


Fig. 1. Schematic cross-sections of the Multiple-Gate devices classified as a function of the "Equivalent Number of Gates" (EGN)

The scaling of Multiple-Gate MOSFET requires the use of an increasingly thinner Silicon film, for which new phenomena have to be taken into account, such as quantum-mechanical confinement. These phenomena induce a strong subband splitting and the carrier confinement in the narrow potential well formed by the Silicon film (Taur & Ning, 1998). Quantum effects sensibly modify the three dimensional (3-D) carrier distribution in the channel, the most important effect being the shift of the charge centroid away from the interfaces into the Silicon film. The inversion charge and then the drain current are reduced in the quantum case with respect to the "classical" case (i.e. without quantum effects). Quantum-mechanical confinement is stronger when the film is thinner. It has been shown that the energy quantization becomes important for channels below 10 nm thick, for which it becomes mandatory to take into account quantum effects in the device simulation (Bescond et al., 2004). In Single-Gate or Double-Gate configurations the carriers are confined in a single direction (vertically, perpendicular to the gate electrode and to the source-to-drain axis). In multiple-gate architectures, and especially in Gate-All-Around devices, the quantum-mechanical confinement is stronger because the carrier energy is quantified in two directions (vertically but also horizontally, in both directions perpendicular to the gates electrodes and to the source-to-drain axis). Then, the carrier confinement and its effects (such as the reduction of the total inversion charge) are stronger in Multiple-Gate devices with EGN≥3 than for single-gate or double-gate architectures.

As the MOSFET is scaling down, the sensitivity of integrated circuits to radiation, coming from the natural space or present in the terrestrial environment, has been found to seriously increase (Baumann, 2005; Dodd, 1996; Dodd & Massengill, 2003; Dodd, 2005). In particular, ultra-scaled memory ICs are more sensitive to single-event-upset (SEU) and digital devices

are more subjected to digital single-event transient (DSETs). Single-event-effects (SEE) are the result of the interaction of highly energetic particles, such as protons, alpha particles, or heavy ions, with sensitive regions of a microelectronic device or circuit. These SEE may perturb the device/circuit operation (e.g., reverse or flip the data state of a memory cell, latch, flip-flop, etc.) or definitively damage the circuit (e.g. gate oxide rupture, destructive latch-up events).

Modeling and simulating the effects of ionizing radiation has long been used for better understanding the radiation effects on the operation of devices and circuits. In the last two decades, due to substantial progress in simulation codes and computer performances which reduce computation times, simulation reached an increased interest. Due to its predictive capability, simulation offers the possibility to reduce radiation experiments and to test hypothetical devices or conditions, which are not feasible (or not easily measurable) by experiments. Physically-based numerical simulation at device-level presently becomes an indispensable tool for the analysis of new phenomena specific to short-channel devices (non-stationary effects, quantum confinement, quantum transport), and for the study of radiation effects in new device architectures (such as multiple-gate, Silicon nanowire MOSFET), for which experimental investigation is still limited. In these cases, numerical simulation is an ideal investigation tool for providing physical insights and predicting the operation of future devices expected for the end of the roadmap. A complete description of the modeling and simulation of SEE, including the history and the evolution of this research domain, have been presented in the reference survey papers by Dodd (Dodd, 1996; Dodd & Massengill, 2003; Dodd, 2005) and Baumann (Baumann, 2005).

In a previous work (Munteanu et al., 2006), we investigated the impact of the quantum effects on the transient response of 50 nm gate length Fully-Depleted (FD) Single-Gate MOSFET with 11 nm thick Silicon film. In that work, we found an excellent agreement between experimental bipolar gain values (measured by heavy ions experiments) and simulated bipolar gain obtained by quantum-mechanical simulation. The results were also consistent with experimental data obtained by pulsed laser irradiation performed on 50 nm gate length transistors fabricated with the same technology (Ferlet-Cavrois et al., 2005). The study presented in (Munteanu et al., 2006) illustrated the importance of taking into account quantum effects in the simulation of the device response when submitted to heavy ion irradiation. The simulation results also showed that even if the impact of quantum effects can be considered as limited in these 11 nm thick FD Single-Gate devices, it will become important for thinner films and for double-gate architectures.

The transient response of Multiple-Gate nanowire MOSFETs under heavy ion irradiation has been already addressed (Castellani et al., 2006; Francis et al., 1995), but to the best of our knowledge, all the previous studies considered the "classical" approach. In this work we use 3D quantum numerical simulation for investigating the drain current transient produced by the ion strike in Multiple-gate nanowire MOSFETs with ultra-thin channels (≤ 10 nm). We firstly consider devices with a gate length of 32 nm and 10 nm-thick Silicon film. For these devices we compare the classical and quantum simulation in terms of drain current transient induced by the ion strike, carrier density and bipolar amplification. Three different Multiple-Gate configurations are considered: Double-Gate, Triple-Gate and Gate-All-Around. In a second step, the devices scaling is addressed and the impact of the quantum effects is analyzed for two cases: (a) 32 nm gate length devices with thinner film (8 nm and 5 nm) and (b) completely scaled devices with 25 nm gate length (8 nm thick film) and 20 nm gate length (5 nm thick film). For each point the classical and the quantum results are

compared and the differences between the four architectures from the view-point of the devices immunity to heavy ion irradiation are analyzed.
This chapter is organized as follows: section 2 presents a detailed description of simulated devices and section 3 describes the simulation code, including the modelling of quantum confinement effects and the simulation of the effects of an ion strike. Section 4 details the simulation of transient effects in Multiple-Gate MOSFET submitted to heavy-ion irradiation. Static and transient characteristics calculated in both quantum and classical cases are presented and compared. Finally, a detailed study concerns the impact of device scaling on the transient response to radiation effects.

2. Simulated devices

In this work, we simulate square cross-section nanowire Double-Gate, Triple-Gate and Gate-All-Around MOSFETs with 32 nm, 25 nm and 20 nm physical gate lengths. The description of the 3-D architectures considered in the simulation and the definition of their geometrical parameters are represented in Fig. 2.

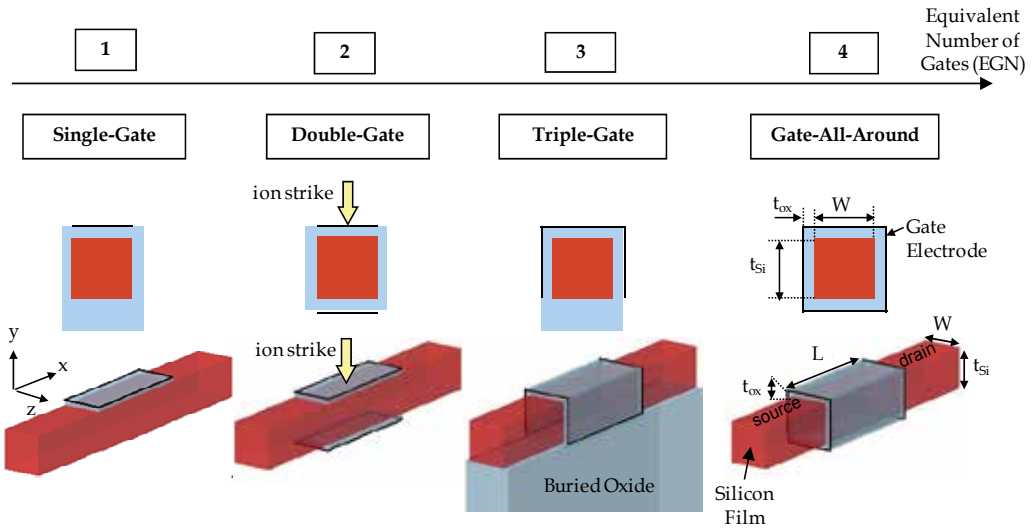


Fig. 2. Schematic description of the 3-D simulated Double-Gate, Triple-Gate and Gate-All-Around structures and their main geometrical parameters considered in this work. The Single-Gate structure is also shown for comparison. The devices are classified as a function of the "Equivalent Number of Gates" (EGN). The schematic cross-sections in the (y-z) plane are also shown. For all the simulated structures, there is no gate overlap with the S/D regions and the S/D doping concentration is $10^{20}cm^{-3}$. The position of the ion strike is also indicated by the arrow; the ion strikes vertically in the middle of the channel (between the source and drain region) and in a direction parallel to the y axis. The Silicon substrate was simulated for the Triple-Gate structures. All structures have Silicon film with square section (t_{Si}=W)

All devices have been calibrated to fill the ITRS requirements for Low Power Technology in terms of drain current in the off-state (I_{OFF}<5×10^{-3} A/μm) (ITRS, 2009). The Silicon film and

the gate oxide have the following dimensions: (a) t_{Si}=W=10 nm and t_{ox}=1.2 nm for devices with L=32 nm gate length, (b) t_{Si}=W=8 nm and t_{ox}=1 nm for devices with L=25 nm and (c) t_{Si}=W=5 nm and t_{ox}=0.9 nm for devices with L=20 nm. All devices have intrinsic channel and mid-gap gate, and the thickness of the buried oxide is 100 nm. The supply voltage is 0.8 V for devices with L=32 nm and L=25 nm and 0.7 V for devices with L=20 nm.

3. Description of the simulation code

3-D numerical simulations have been performed with both 3-D Sentaurus code (Sentaurus, 2009) and with our full quantum homemade Fortran code BALMOS3D (Munteanu & Autran, 2003). The physical models considered in the Sentaurus code include the SRH and Auger recombination models and the Fermi-Dirac carrier statistics.
Concerning the transport modelling, the drift-diffusion (DD) model was for many years the standard level of solid-state device modelling, mainly due to its simple concept and short simulation times. This approach is appropriate for devices with large feature lengths. This model considers that carrier energy does not exceed the thermal energy and carrier mobility is only a local function of the electric field (mobility does not depend on carrier energy). These assumptions are acceptable as long as the electric field changes slowly in the active area, as is the case for long devices (Munteanu & Autran, 2008). When the device feature size is reduced, the electronic transport becomes qualitatively different from the DD model since the average carrier velocity does not depend on the local electric field. In short devices steep variations of electric field take place in the active area of the devices. Then, non-stationary phenomena occur following these rapid spatial or temporal changes of high electric fields. Since these phenomena play an important role in small devices, new advanced transport models become mandatory for accurate transport simulation. The hydrodynamic model, obtained by taking the first three moments of the Boltzmann Transport Equation (BTE), represents the carrier transport effects in short devices more accurately than the DD model. The hydrodynamic model is a macroscopic approximation to the BTE taking into account the relaxation effects of energy and momentum. This model removes several limiting assumptions of DD: the carrier energy can exceed the thermal energy and all physical parameters are energy-dependent (Munteanu & Autran, 2008). In this work we use the hydrodynamic model for the transport modelling. Then, both the impact ionization and the carrier mobility depend on carrier energy calculated with the hydrodynamic model. The mobility model also includes the dependence on the lattice temperature and on the channel doping level. The mobility also depends on the doping level and the lattice temperature. Quantum confinement effects have been considered in the simulation using the Density Gradient model, as explained in section 3.1.

3.1 Modeling of quantum effects

The aggressive scaling-down of bulk MOSFETs in the deep submicrometer domain requires ultrathin oxides and high channel doping levels for minimizing the drastic increase of short channel effects. The direct consequence is a strong increase of the electric field at Si/SiO_2 interface, which creates a sufficiently steep potential well for inducing the quantization of carrier energy (Munteanu & Autran, 2008). In bulk architecture, carriers are then confined in a vertical direction in a quantum well (formed by the Silicon conduction band bending at the interface and the oxide/Silicon conduction band-offset) having feature size close to the electron wavelength. This gives rise to a splitting of the energy levels into subbands (two-

dimensional (2-D) density of states) (Hareland et al., 1998), such that the lowest of the allowed energy levels for electrons (resp. for holes) in the well does not coincide with the bottom of the conduction band (resp. the top of the valence band). In addition, the total density of states in a 2-D system is less than that in a three-dimensional (3-D) (or classical) system, especially for low energies. Carriers occupying the lowest energy levels behave like quantized carriers while those lying at higher energies, which are not as tightly confined in the potential well, can behave like classical (3-D) particles with three degrees of freedom (Munteanu & Autran, 2008). As the surface electric field increases, the system becomes more quantized as more and more carriers become confined in the potential well. The quantum-mechanical confinement considerably modifies the carrier distribution in the channel: the maximum of the inversion charge is shifted away from the interface into the Silicon film (Munteanu & Autran, 2008). Because of the smaller density of states in the 2-D system, the total population of carriers will be smaller for the same Fermi level than in the corresponding 3-D (or classical) case. This phenomenon affects the net sheet charge of carriers in the inversion layer, thus requiring a larger gate voltage in order to populate a 2-D inversion layer to have the same number of carriers as the corresponding 3-D system. This leads to an increase of the threshold voltage of a MOSFET, which is an important issue, especially as the power supply voltages drop to lower levels. The gate capacitance and carrier mobility are also modified by quantum effects. These considerations indicate that the wave nature of electrons and holes can no longer be neglected in ultra-short devices and have to be considered in simulation studies. Quantum confinement becomes also important for the device response to single events.

Various methods have been suggested to model these quantum effects. Among the approaches that are compatible with classical device simulators based on the drift-diffusion (or hydrodynamic) approach, the physically most accurate method is to include the Schrödinger equation into the self-consistent computation of the device characteristics (Stern, 1972). However, solving the Schrödinger equation in itself is very much time-consuming. Various simpler methods have been suggested, such as the Van Dort model or the Hansch model. The van Dort model (van Dort, 1994) expresses the quantum effect by an apparent band edge shift that is a simple function of the electric field. The model is based on the expression for the lowest eigenenergy of a particle in a triangular potential and reproduces the characteristics obtained with the Schrödinger equation quite well. However, this model does not give the correct charge distribution in the device. The Hansch (Hansch et al., 1989] model proposes a quantum correction of the density of states as a function of depth below the Si/SiO_2 interface. The charge distribution is better reproduced, but the model strongly overestimates the impact of quantum effects on the drain current characteristics.

Other alternative to take into account quantum confinement of carriers is the Density-Gradient model (Ancona & Iafrate, 1989; Grubin et al., 1993; Wettstein & al., 2002), coupled with the Drift-Diffusion or the hydrodynamic transport equations. The Density-Gradient model considers a modified equation of the electronic density including an additional term dependent on the gradient of the carrier density. To include quantization effects in a classical device simulation, a simple approach is to introduce an additional potential-like quantity Λ in the classical electron density formula, as follows (Sentaurus, 2009):

$$n = N_C \exp\left(\frac{E_{Fn} - E_C - \Lambda}{kT}\right) \tag{1}$$

where n is the electron density, T is the carrier temperature, k is the Boltzmann constant, N_C is the conduction band density of states, E_C is the conduction band energy, and E_{Fn} is the electron Fermi energy. The impact of the quantum confinement on the carrier density in the device can be taken into account by properly modelling the quantity Λ. For the Density Gradient model, Λ is given in terms of a partial differential equation:

$$\Lambda = -\frac{\gamma\hbar}{6m}\frac{\nabla^2\sqrt{n}}{\sqrt{n}} \tag{2}$$

where $\hbar = h/2\pi$ is the reduced Planck constant, m is the density of states mass, and γ is a fit factor. An equation similar to (1) applies for the holes density. These new equations for electrons and holes density are then used in the self-consistent resolution of the Poisson equation and of the transport equation (Drift-Diffusion or hydrodynamic model), as explained in (Munteanu & Autran, 2008).

3.2 Calibration of the simulation code

It has been shown that the Density Gradient model can accurately account for quantum carrier confinement in Single-Gate SOI and Double-Gate devices with an appropriate calibration step of the fit factor γ (Wettstein & al., 2002). In this work, we have used the exact solution of the Schrödinger –Poisson system of equations (as given by BALMOS3D) for calibrating the Density Gradient model.

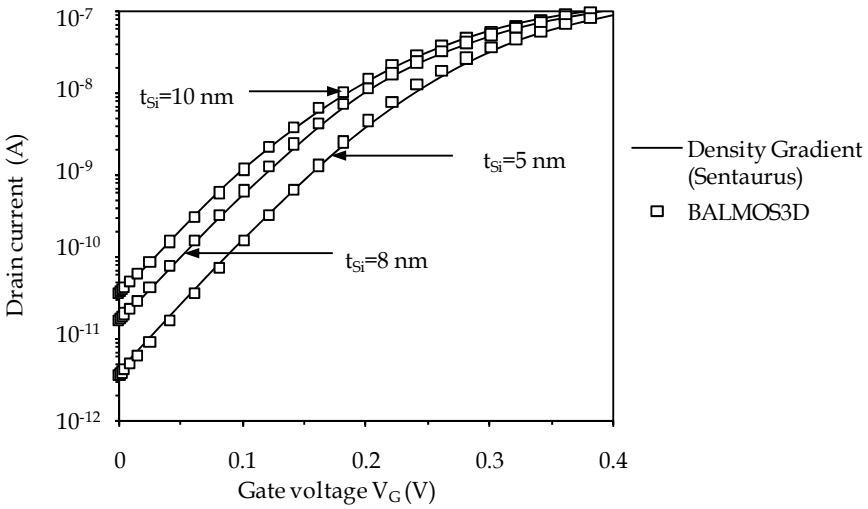


Fig. 3. Calibration of the Density-Gradient model (Sentaurus, 2009) on BALMOS3D. The simulated devices are 32 nm gate length Double-Gate MOSFETs with three Silicon thicknesses (t_{Si}=10 nm, 8 nm and 5 nm). For better illustration, the figure only shows the subthreshold region of the drain current characteristics. V_D=0.8 V

BALMOS3D (Munteanu & Autran, 2003) is a homemade full quantum Fortran simulator, which solves self-consistently the Schrödinger equation and the Poisson equations on a 3-D-grid. The solution of this system of equation is coupled with the Drift-Diffusion transport equation in the channel. A finite difference scheme with a non-uniform mesh has been

considered on a 3-D domain, which includes the channel, the source and drain regions, the gate oxide layers and the gate electrodes. Electric field penetration in the source/drain and electron wave-function penetration in the gate oxide can be also taken into account.
A calibration step of the Density Gradient model on BALMOS3D has been performed on each simulated device, for obtaining the fit factor γ. This factor has different values as a function of the film thickness and gate length. For each particular device, the drain current static characteristics as a function of the gate bias, $I_D(V_G)$, has been computed with BALMOS3D. The same device (with identical geometry) has been implemented in the 3D Sentaurus code and its $I_D(V_G)$ characteristic has been simulated, taken into account the Density Gradient model. The fit factor γ has been then finely tuned in order to obtain a perfect match between the characteristics calculated with BALMOS3D and that simulated by Sentaurus. Figure 3 shows an example of the calibration step on 32 nm gate length Double-Gate MOSFET with three different Silicon film thicknesses.

3.3 Modeling the effect of a particle strike

The physical parameters calibrated previously have been further used in the simulation of drain current transients produced by an ion strike on the sensitive regions of the device. The drain current transients have been simulated in two cases: the classical case (i.e. without quantum effects) and in the quantum case (using Density Gradient model with the fit factor γ as calibrated on BALMOS3D).
The radiation effects have been simulated using the HeavyIon module (Sentaurus, 2009), considering an electron-hole pair column centred on the ion track axis to model the ion strike. The ion track structure to be used as input in simulation is presently a major issue for device simulation. The first representations included a simple cylindrical charge generation with a uniform charge distribution and a constant LET along the ion path. However, the real ion track structure is radial and varies as the particle passes through the matter. When the particle strikes a device, highly energetic primary electrons (called δ-rays) are released. They further generate a very large density of electron-hole pairs in a very short time and a very small volume around the ion trajectory, referred as the ion track. These carriers are collected by both drift and diffusion mechanisms, and are also recombined by different mechanisms of direct recombination (radiative, Auger) in the very dense core track, which strongly reduces the peak carrier concentration. All these mechanisms modify the track distribution both in time and space. As the particle travel through the matter, it loses energy and then the δ-rays become less energetic and the electron-hole pairs are generated closer to the ion path. Then, the incident particle generates characteristic cone-shaped charge plasma in the device (Dodd, 2005).
The real ion track structure has been calculated using Monte-Carlo methods (Hamm et al., 1979; Martin et al., 1987; Oldiges et al., 2000). These simulations highlighted important differences between the track structure of low-energy and high-energy particles, even if the LET is the same (for details see (Dodd et al., 1998; Dodd, 2005)). High-energy particles are representative for ions existing in the real space environment, but they are not available in typical laboratory SEU measurements (Dodd, 1996). Then the investigation of the effects of high-energy particles by simulation represents an interesting opportunity, which may be difficult to achieve experimentally.
Analytical models for ion track structure have been also proposed in the literature and implemented in simulation codes. One of the most interesting models is the “non-uniform

power law" track model, based on the Katz theory (Kobetich & Katz, 1968) and developed by Stapor (Stapor & McDonald, 1988). In this model, the ion track has a radial distribution of excess carriers expressed by a power law distribution and allows the charge density to vary along the track (Dussault et al., 1993). Other analytical models propose constant radius non-uniform track or Gaussian distribution non-uniform track.
In commercial simulation codes, the effect of a particle strike is taken into account as an external generation source of carriers. The electron-hole pair generation induced by the particle strike is included in the continuity equations via an additional generation rate. This radiation-induced generation rate can be connected to the parameters of irradiation, such as the particle Linear Energy Transfer (LET). The LET is the energy lost by unit of length (-dE/dl), which is expressed here in MeV cm²/mg (1pC/μm≈100MeV cm²/mg). The particle LET can be converted into an equivalent number of electron-hole pairs by unit of length using the mean energy necessary to create an electron-hole pair (E_{ehp}) (Roche, 1999):

$$\frac{dN_{ehp}}{dl} = \frac{1}{E_{ehp}} \frac{dE}{dl} \tag{3}$$

where N_{ehp} is the number of electron-hole pairs created by the particle strike. By associating two functions describing the radial and temporal distributions of the created electron-hole pairs, the number of electron-hole pairs is included in the continuity equations (Munteanu & Autran, 2008) via the following radiation-induced generation rate:

$$G(w,l,t) = \frac{dN_{ehp}}{dl}(l) \cdot R(w) \cdot T(t) \tag{4}$$

where R(w) and T(t) are the functions of radial and temporal distributions of the radiation induced pairs, respectively. Equation (4) assumes the following hypothesis: the radial distribution function R(w) depends only on the distance traversed by the particle in the material and the generation of pairs along the ion path follows the same temporal distribution function in any point. Since function G must fill the condition:

$$\int_{w=0}^{\infty} \int_{\theta=0}^{2\pi} \int_{t=-\infty}^{\infty} G\, w\, dw\, d\theta\, dt = \frac{dN_{ehp}}{dl} \tag{5}$$

functions R(w) and T(t) are submitted to the following normalization conditions:

$$2\pi \int_{w=0}^{\infty} R(w)\, w\, dw = 1 \tag{6}$$

$$\int_{t=-\infty}^{\infty} T(t)\, dt = 1 \tag{7}$$

The ion track models available in commercial simulation codes usually propose a Gaussian function for the temporal distribution function T(t):

$$T(t)=\frac{e^{-(\frac{t}{t_C})^2}}{t_C\sqrt{\pi}} \tag{8}$$

where t_C is the characteristic time of the Gaussian function which allows one to adjust the pulse duration. The radial distribution function is usually modelled by an exponential function or by a Gaussian function:

$$R(w)=\frac{e^{-(\frac{w}{r_C})^2}}{\pi r_C^2} \tag{9}$$

where r_C is the characteristic radius of the Gaussian function used to adjust the ion track width. Previous works have demonstrated that the different charge generation distributions used for the radial ion track does affect the device transient response, but the variation is typically limited to ~5% for ion strikes on bulk p-n diodes (Dodd, 2005; Dussault et al., 1993). Considering a LET which is not constant with depth along the path has a more significant impact on the transient response in bulk devices. The key parameters of the single event transient (peak current, time to peak and collected charge) have up to 20% variation when LET is allowed to vary with depth compared to the case of a constant LET (Dussault et al., 1993). Nevertheless, the LET variation with depth has no influence on the transient response of actual SOI devices with thin Silicon film.
In this work, the irradiation track simulated in vertical incidence has a Gaussian shape with narrow radius (14 nm) and a Gaussian time dependence, centred on 10 ps and with a characteristic width of 2 ps. The ion strikes in the middle of the channel. The deposited charge is calculated considering the Gaussian distribution of the ion track and the 3D geometry of the Silicon film. The collected charge is given by the integration of the drain current over the transient duration and the bipolar amplification is finally calculated as the ratio between the collected and deposited charges, as it will be shown in section 4.3.

4. Simulation of multiple-gate devices

4.1 Static characteristics

Figure 4 shows the quantum confinement directions in three different generic configurations: Single-Gate, Double-Gate and Gate-All-Around devices. The impact of quantum effects on the electron density extracted along a cut-line parallel to the confinement directions is also illustrated for the three devices.
In the Single-Gate devices carriers are confined in a very narrow triangular potential well, formed at the Si/SiO_2 interface. The quantum carrier density in the y direction is then modified as compared to the classical one: the classical electron density is maximal at the Si/SiO_2 interface, since the quantum density profile show a maximum shifted inside the Silicon film at several nanometers depth. Then, the electron density near the interface (as well as the total electron charge in the conduction channel) is strongly reduced. In the case of a Double-Gate archtecture, the potential well is rectangular and its dimension is now controlled by the Silicon film thickness, which becomes a key parameter in the quantum effects analysis. Similar to the Single-Gate configuration, the electron density is maximum at

the two interfaces in the classical case. In the quantum case, the density profile has two maxima situated within the Silicon film at several nanometers depth from each interface. Our results are in perfect concordance with (Majkusiak et al., 2002), where quantum effects are simulated using the solution of the 1-D Schrödinger equation. The drain current is splitted in two separate channels, but they are no more located at the interface as in the classical case. Finally, in the Gate-All-Around structure, carriers are confined in a double rectangular potential well (along the y and the z directions), which considerably enhances the quantum confinement effects. The carrier motion is no more free in the z direction (as is the case of the Single-Gate and the Double-Gate devices), but their energy is quantized as in the y direction. Both the gate electrode width (W) and the Silicon film thickness control here the quantum effects. The quantum electron density in the z direction is no more maximal at the interface but has two maxima moved into the Silicon film as for the carrier density in the y direction. Then the total inversion charge is lower than in the Double-Gate configuration.

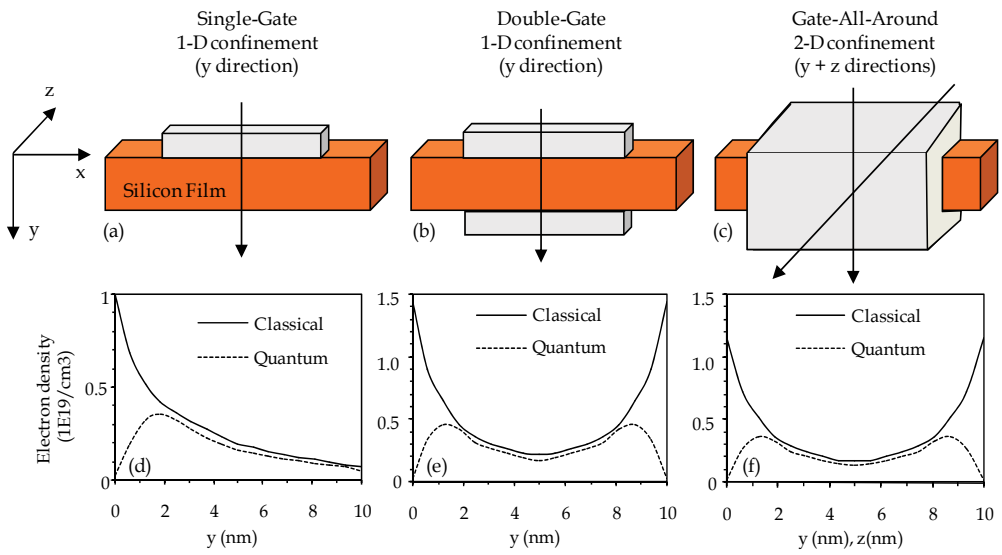


Fig. 4. Schematically representation of the quantum-mechanical confinement directions in (a) Single-Gate, (b) Double-Gate and (c) Gate-All-Around configurations. The profile of the carrier density in a cut-line along the film thickness is also reported for both classical and quantum cases: (d) Single-Gate, (e) Double-Gate and (f) Gate-All-Around. V_D=V_G=0.8 V, L=32 nm

The $I_D(V_G)$ curves for the different 32 nm Multiple-Gate MOSFET architectures simulated in the classical and quantum cases are shown in Fig. 5.

The results show that increasing the "equivalent number of gates" reduces the off-state current (Munteanu et al., 2007) and improves the subthreshold swing S (S = 70 mV/dec for Double-Gate, S = 68.5 mV/dec for Triple-Gate and S = 61.5 mV/dec for Gate-All-Around). This is due to the better electrostatic control of the gate over the channel that reduces short channel effects. At the same time, the on-state current increases with EGN (Fig. 5), due to the multiple-channel conduction. As expected, the quantum current is lower than the classical

one, because the total inversion charge is reduced in the quantum case. Figure 5 also shows that the difference between the classical and the quantum off-state current increases when going from Double-Gate to Gate-All-Around device. The ratio between the classical and quantum off-state currents is reported in Table 1 for the three considered configurations.

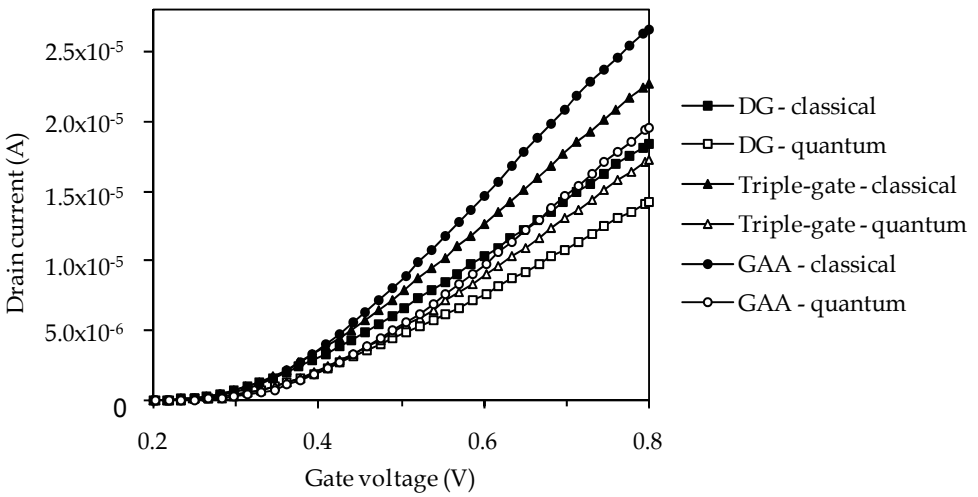


Fig. 5. Drain current $I_D(V_G)$ characteristics in classical and quantum-mechanical cases for 32 nm Double-Gate, Triple-gate, Ω-Gate and Gate-All-Around architectures (V_D=0.8 V). The quantum drain current was simulated using the Density-Gradient model calibrated on BALMOS3D numerical results

	t_{Si}=W=10 nm L=32 nm	t_{Si}=W=8 nm L=25 nm	t_{Si}=W=5 nm L=22 nm
Double-Gate (EGN=2)	1.91	2.02	3.01
Triple-Gate (EGN=3)	2.24	2.3	3.66
Gate-All-Around (EGN=4)	2.36	2.67	4.11

Table 1. Ratio Ioff_cl/Ioff_q of the off-state currents in classical (Ioff_cl) and quantum (Ioff_q) approaches for the three technological nodes studied in this work. The quantum drain current has been calculated using the Density Gradient model calibrated on BALMOS3D for each configuration.

We remark that this ratio increases with EGN for a given technology node. This effect can be explained by the dimensionality of the confinement. In Double-Gate, carriers are confined in one direction (y direction), since in Triple-Gate and Gate-All-Around carriers are confined in two directions (y and z), which strongly enhances the energy quantization with respect to the Double-Gate case.

4.2 Transient simulation results

The time evolution of the electron density distribution in a vertical cross-section (y-z plane) in the middle of the channel is represented in Fig. 6 for three configurations: Double-Gate, Tri-Gate and Gate-All-Around. We observe that for all devices the quantum electron charge is centred in the middle of the film and the electron density has lower values than in the classical case. In off-state bias condition, the carrier conduction in all devices is mainly dominated by the volume inversion phenomenon: carriers flow from source to drain over the entire Silicon film thickness. In consequence, the off-state current is directly proportional to the film thickness. In the quantum case the volume inversion phenomenon is reinforced because the quantum carrier density becomes more centred in the middle of the film (Fig. 6). This effect is enhanced when EGN increases from 2 (Double-Gate) to 4 (Gate-All-Around), as illustrated in Fig. 6.

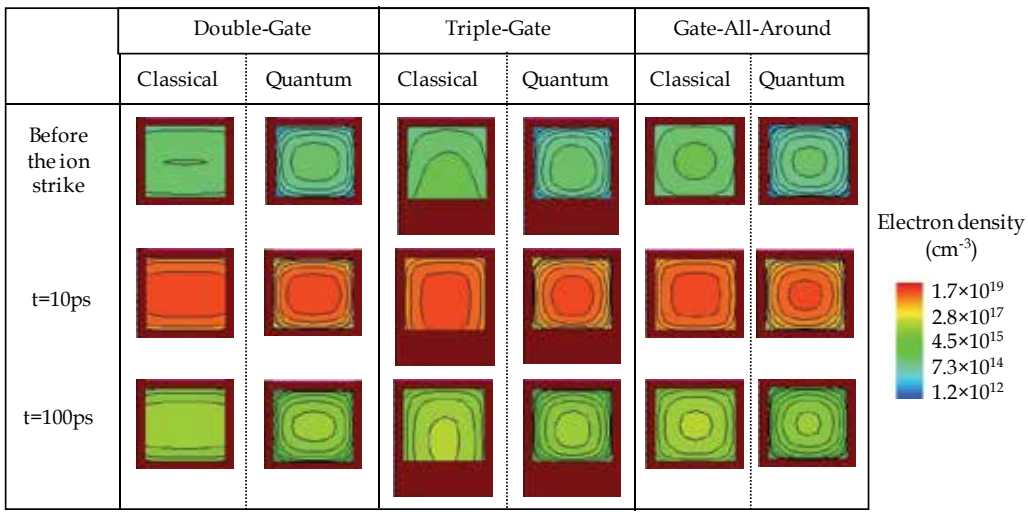


Fig. 6. Classical and quantum electron density (expressed in cm^{-3}) in a vertical cross-section (y-z plane) in the middle of the channel of 32 nm Double-Gate, Triple-Gate and Gate-All-Around at different times before and after the ion strike. The devices are biased in the off-state at V_G=0 V and V_D=0.8 V. The brown regions represent the gate oxide (in Double-Gate and Gate-All-Around devices) and the gate and buried oxide in Triple-Gate devices

The drain current transients produced by the ion strike are illustrated in Fig. 7 for the classical case and for a LET value of 1 MeV/(mg/cm^2). The four configurations corresponding to the 32 nm gate length ITRS LP technology node are simulated in the off-state. The peak value of the drain current transient is reduced when EGN increases. When EGN increases, the channel is better controlled by the gate and the floating body effects are strongly reduced. Then the drain current transient tail is shorter when going from Double-Gate to Gate-All-Around devices. Figure 8 compares the classical and the quantum drain current transient for two configurations: Double-Gate and Gate-All-Around devices with 32 nm gate length. As expected, the peak of the quantum drain current transient is lower than the classical one for both configurations, due to the quantum confinement which induces lower quantum off-state current.

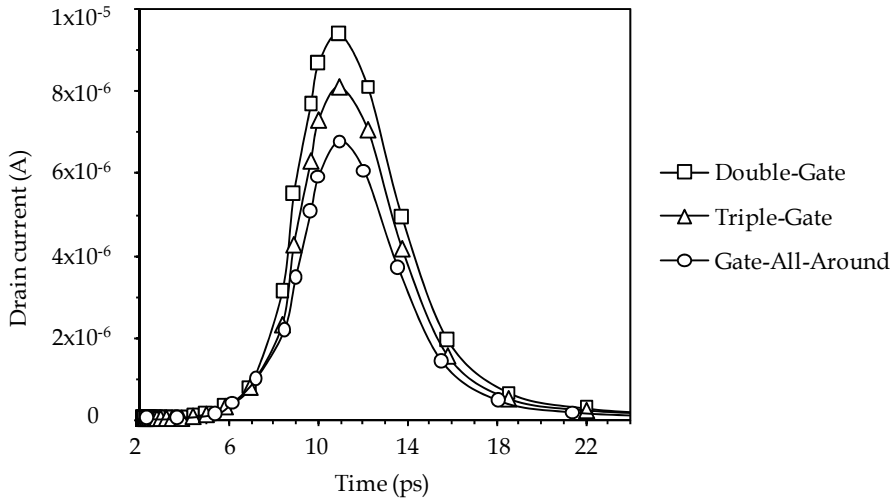


Fig. 7. Drain current transients induced by an ion strike vertically (y direction) in the middle of the Silicon film (classical simulation). The ion track generation has a Gaussian shape versus time (characteristic time of 2 ps), centred at 10 ps and a LET=1 MeV/(mg/cm^2). The simulated devices are 32 nm gate length MOSFETs. All devices are off-state biased (V_G=0 V, V_D=0.8 V)

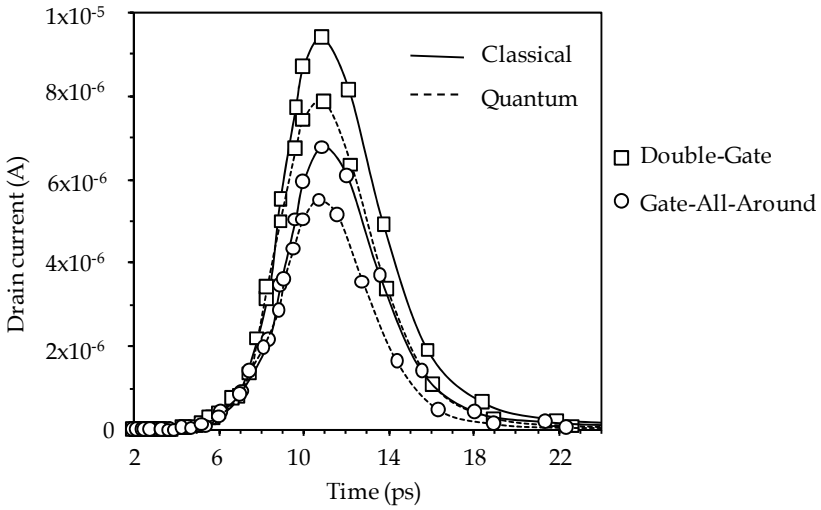


Fig. 8. Drain current transients induced by an ion strike vertically (y direction) in the middle of the Silicon film. Comparison between classical and quantum simulation in Double-Gate and Gate-All-Around MOSFETs. All devices are off-state biased (V_G=0 V, V_D=0.8 V)

4.3 Bipolar amplification

The bipolar amplification is a phenomenon specific to partially-depleted SOI devices and its basic mechanism was largely explained and simulated in previous works (Ferlet-Cavrois et

al., 2002; Ferlet-Cavrois et al., 2004; Schwank et al., 2003). Bipolar amplification can also occur in fully depleted devices, as those studied here.
The bipolar transistor mechanism in fully depleted devices has been explained in (Brisset et al., 1994) using Monte Carlo simulations of 0.25 μm fully depleted SOI transistors: after irradiation of a n-channel MOSFET biased in its off state, excess holes are accumulated in the channel (mainly near the gate oxide) and lower the potential barrier; then electrons diffuse from source to drain to maintain the electrical neutrality. This mechanism is comparable to the bipolar transistor effect in partially depleted SOI transistors (Massengill et al., 1990). Because bipolar amplification is less important for fully depleted than for partially depleted devices, circuits based on fully depleted transistors are less sensitive to single-event upset than partially depleted circuits (Ferlet-Cavrois et al., 2002).
The effect of the parasitic bipolar transistor in SOI devices is quantified using the bipolar gain, β. The bipolar gain corresponds to the amplification of the deposited charge and is given by the ratio between the total collected charge, Q_{coll}, at the drain electrode and the deposited charge, Q_{dep}:

$$\beta = \frac{Q_{coll}}{Q_{dep}} \tag{10}$$

The total collected charge at drain electrode is given by:

$$Q_{coll} = \int_0^{\infty} I_D dt \tag{11}$$

The deposited charge in a SOI device is calculated as a function of the particle LET using the following equation (Ferlet-Cavrois, 2004; Munteanu & Autran, 2008):

$$Q_{dep}[fC] = 10.3 \times LET[MeV/(mg/cm^2)] \times t_{Si}[\mu m] \tag{12}$$

where t_{Si} is the Silicon film thickness and 10.3 is a multiplication factor for Silicon calculated using the Silicon density and the energy needed for creating an electron-hole pair in Silicon (– 3.6 eV) (Ferlet-Cavrois, 2004; Munteanu & Autran, 2008). In this equation a normal incident ion strike is considered and the LET is supposed constant along the ion path in the active Silicon film.
The bipolar gain for 32 nm gate length Multiple-Gate devices in both classical and quantum cases is shown in Fig. 9 as a function of the LET value. The bipolar amplification decreases when increasing EGN due to less floating body effects. However, at high LET (>2 MeV/(mg/cm²)), the classical bipolar gain becomes the same for all configurations. This can be explained by the huge deposited charge by the ion which masks the impact of other phenomena such as the electrostatic control by the gate.
Previous experimental and theoretical studies showed that, generally, fully depleted SOI-based devices (with either single- or double-gate configuration) present reduced floating body effects and then lower bipolar amplification of the collected charge than partially-depleted SOI devices (Ferlet-Cavrois et al., 2002; Ferlet-Cavrois et al., 2005). In Multiple-Gate devices the control of the channel by the gates is naturally reinforced, and reduces even more the floating body effects. Then very low values are obtained for the bipolar gain. Our results are consistent with simulation data from (Munteanu et al., 2006) and (Castellani et

al., 2006), but they are very low compared with those expected by extrapolation from simulations in (Dodd et al., 2004). This is probably due to the partially depleted SOI Single-Gate structures used in (Dodd et al., 2004), whereas ultra-thin fully-depleted devices and multiple-gate configurations are considered here.
The quantum bipolar gain is lower than the classical one, excepted at very high LET (Fig. 9). Our results show that two phenomena, with opposite effects on the bipolar gain, are to be considered. On one hand, the lower off-state current in the quantum case leads to a lower quantum bipolar amplification (Castellani et al., 2006). On the other hand, in the quantum case, the electron density is lower leading to slower recombination process (reflected in a longer transient tail) and then to a higher collected charge. Depending on the injection regime, one phenomenon or the other prevails. At low injection regime, the generated charge is not very high and carriers recombine rapidly. Then the bipolar gain follows the off-state current behaviour, both being lower for a quantum approach than for a classical one. In very high injection conditions, the electron charge in the film is not sufficient to recombine the enormous generated charge and then, the recombination process is sensibly slower. This has been verified by simulation: the recombination rate in the Silicon film is higher in the classical case than in the quantum case. As a consequence, the quantum collected charge and the quantum bipolar amplification are higher than in the classical ones.

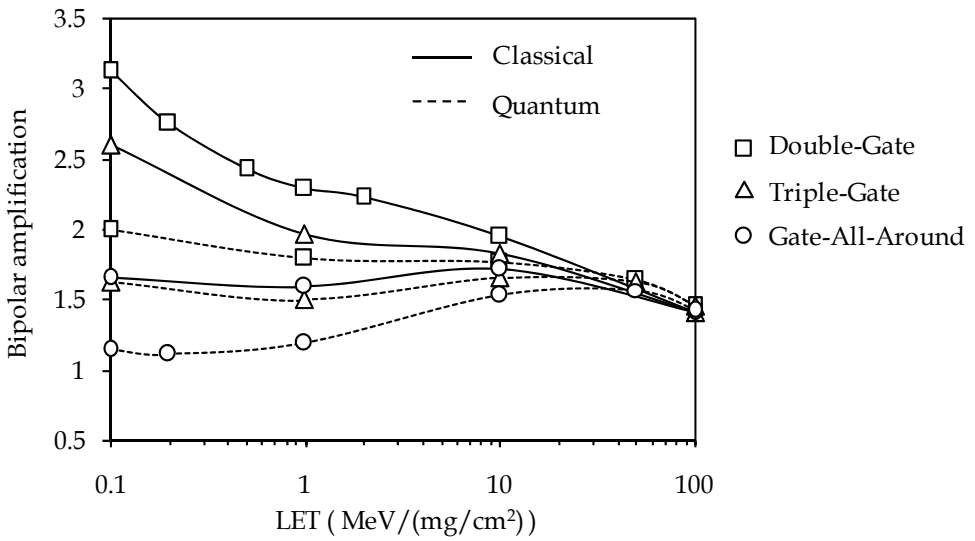


Fig. 9. Simulated classical and quantum bipolar gain as a function of LET in 32 nm gate length Multiple-Gate MOSFET. The transistors are biased in the off-state at V_G=0 V and V_D=0.8 V

5. Device scaling

The effects of the carrier confinement become more important when the Silicon film is thinned because the energy subband splitting is directly proportional with the reverse of the square of the potential well dimension (equal to the film thickness). The ratio between the

classical and quantum off-state currents, reported in Table 1 as a function of t_{Si}, confirms that the quantum confinement is strongly enhanced when the film is thinned down. The collected charge and the bipolar gain (shown in Figs. 10(a) and 10(b)) are lower for thinner channel, in both quantum and classical cases, because the off-state current decreases with the film thickness. The maximal value of the gain is shifted to higher LET values when Silicon film thickness decreases. Our results also indicate that, in the quantum approach, the difference in the bipolar gain when reducing the film thickness (at the same LET) is lower than in the classical case.

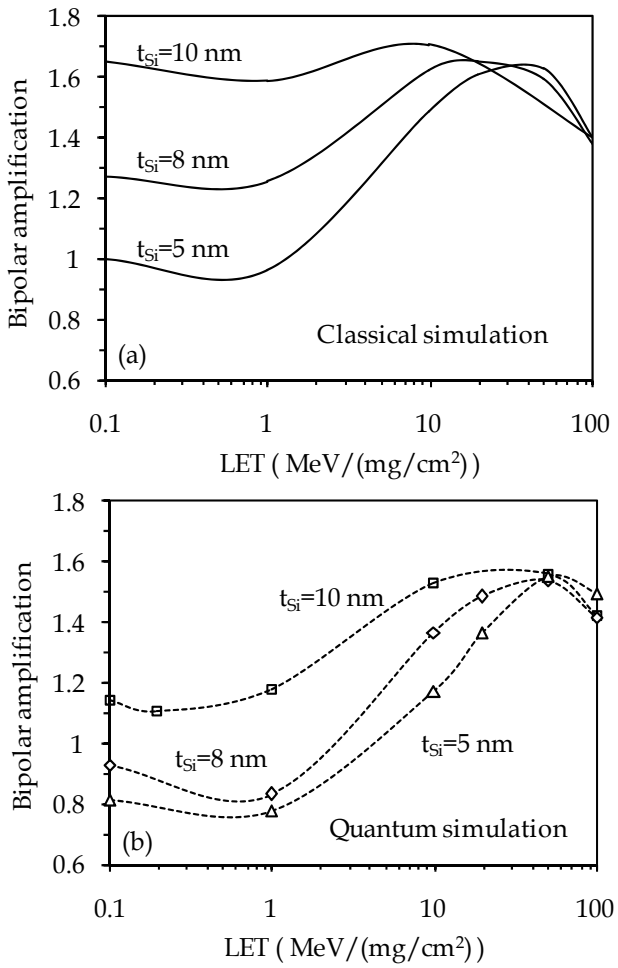


Fig. 10. Bipolar gain variation when reducing the Silicon film thickness in Gate-All-Around MOSFET with 32 nm gate length (the gate width is W=10 nm): (a) classical simulation; (b) quantum simulation. The transistors are biased in the off-state at V_G=0 V and V_D=0.8 V

The quantum bipolar gain for Multiple-Gate devices scaled down to 20 nm gate length and 5 nm Silicon film cross-section was also predicted. As shown in Fig. 11, the difference between the three architectures is reduced for devices with 20 nm gate lengths compared to those with 32 nm and 25 nm gate lengths, due to the very thin square wire cross-section (t_{Si}=W=5

nm). When decreasing the cross-section, the influence of the gate configuration is attenuated and the values of the bipolar gain for the different structures are almost the same. This behaviour can be explained by the fact that, around 5 nm and below, the combination of gate electrostatic control and quantum-mechanical confinement leads to similar carrier density distributions in the film for all gate configurations (Bescond et al., 2004). At this ultimate scale of integration, it should be expected that the sensitivity of all Multiple-Gate nanowire architectures (EGN ≥ 2) to heavy ion irradiation sensibly become equivalent.

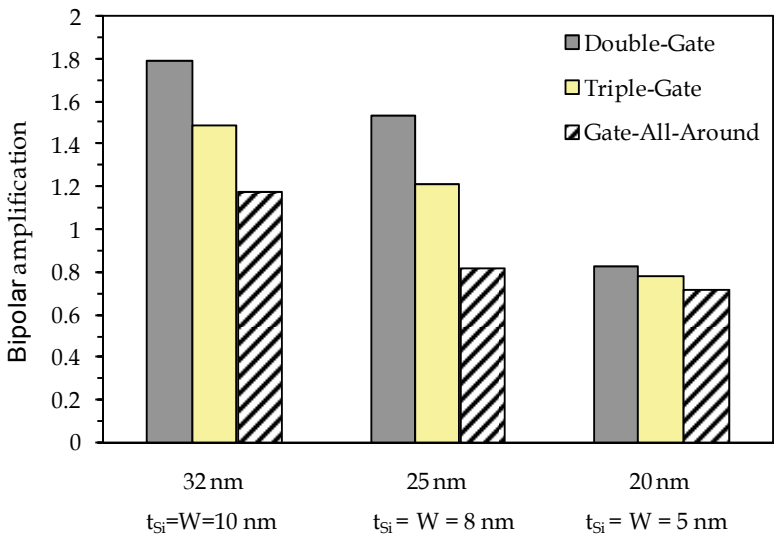


Fig. 11. Bipolar gain calculated in the quantum case for Multiple-Gate nanowire MOSFETs with different gate lengths. The dimensions of the Silicon film cross-section are also indicated. The ion strikes vertically (parallel to the y direction) in the middle of the film

6. Conclusion

In this work we analyzed the impact of quantum confinement on single-event transient immunity of several Multiple-Gate architectures. We showed that the 3-D carrier distribution is strongly affected by the quantum effects, which not only reduces the drain current but also modifies the recombination rate and the charge collection compared to the classical case. Increasing the "number of equivalent gates" induces less floating body effects and then lowers the bipolar gain. Our simulations also showed that when the Silicon channel cross-section is thinned down (around 5nm and below), the bipolar amplification of Multiple-Gate nanowire architectures (EGN ≥ 2) sensibly becomes the same mainly due to carrier quantum confinement.

7. References

Ancona, M.G. & Iafrate, G.J. (1989). Quantum correction to the equation of state of an electron gas in a semiconductor. *Physical Review B*, Vol. 39, No. 13, (May 1989) pp. 9536-9540.

Baumann, R. C. (2005). Radiation-Induced Soft Errors in Advanced Semiconductor Technologies. *IEEE Transactions on Device Material Reliability*, Vol. 5, no. 3, (Sept. 2005) pp. 305-316.

Bescond, M.; Nehari, K.; Autran, J.L.; Cavassilas, N.; Munteanu, D. & Lannoo, M. (2004). 3D Quantum-Modeling and Simulation of Multi-Gate Nanowire MOSFETs. *Proceedings of IEDM Technical Digest*, pp. 617-620, Washington, USA, Dec. 2004, IEEE.

Brisset, C.; Dollfus, P.; Musseau, O.; Leray, J. L. & Hesto, P. (1994). Theoretical study of SEU's in 0.25-pm fully-depleted CMOS/SOI technology. *IEEE Transactions on Nuclear Science*, Vol. 41, No. 6, (1994) pp. 2297-2303.

Castellani, K.; Munteanu, D.; Autran, J.L.; Ferlet-Cavrois, V.; Paillet, P. & Baggio, J. (2006). Investigation of 30 nm Gate-All-Around MOSFET Sensitivity to Heavy Ions: a 3-D Simulation Study. *IEEE Transactions on Nuclear Science*, Vol. 53, No. 4, (Aug. 2006) pp. 1950-1958.

Choi, Y.; Lindert, N.; Xuan, P.; Tang, S.; Ha, D.; Anderson, E.; King, T.; Bokor, J. & Hu, C. (2001). Sub-20nm CMOS FinFET technologies. *Proceedings of IEDM Technical Digest*, pp. 421-424, Washington, USA, Dec. 2001, IEEE.

Colinge, J. P.; Gao, M. H.; Romano-Rodríguez, A.; Maes, H. & Claeys, C. (1990). Silicon-on-insulator "Gate-all-around device". *Proceedings of IEDM Technical Digest*, pp. 595-598, Washington, USA, Dec. 1990, IEEE.

Dodd, P. E. (1996). Device Simulation of Charge Collection and Single-Event Upset. *IEEE Transactions on Nuclear Science*, Vol. 43, No. 2, (1996) pp. 561-575.

Dodd, P. E.; Musseau, O.; Shaneyfelt, M. R.; Sexton, F. W.; D'hose, C.; Hash, G.L.; Martinez, M.; Loemker, R. A.; Leray, J.-L. & Winokur, P. S. (1998). Impact of ion energy on single-event upset. *IEEE Transactions on Nuclear Science*, Vol. 45, No. 6, (Dec. 1998) pp. 2483–2491.

Dodd, P.E. & Massengill, L.W. (2003). Basic mechanisms and modeling of single-event upset in digital microelectronics. *IEEE Transactions on Nuclear Science*, Vol. 50, No. 3, (Jun. 2003) pp. 583–602.

Dodd, P. E.; Shaneyfelt, M. R.; Felix, J. A. & Schwank, J. R. (2004). Production and propagation of single-event transients in high-speed digital logic ICs. *IEEE Transactions on Nuclear Science*, Vol. 51, No. 6, (Dec. 2004) p. 3278-3284.

Dodd, P. E. (2005). Physics-Based Simulation of Single-Event Effects. *IEEE Transactions on Device Material Reliability*, Vol. 5, No. 3, (Sept. 2005) pp. 343-357.

Dussault, H.; Howard, Jr., J. W.; Block, R.C.; Pinto, M. R.; Stapor, W. J. & Knudson, A. R. (1993). Numerical simulation of heavy ion charge generation and collection dynamics. *IEEE Transactions on Nuclear Science*, Vol. 40, No. 6, (Dec. 1993) pp. 1926–1934.

Ferlet-Cavrois, V.; Gasiot, G.; Marcandella, C.; D'Hose, C.; Flament, O.; Faynot, O.; du Port de Pontcharra, J. & Raynaud, C. (2002). Insights on the Transient Response of Fully and Partially Depleted SOI Technologies Under Heavy-Ion and Dose-Rate Irradiations. *IEEE Transactions on Nuclear Science*, Vol. 49, No.6, (Dec. 2002) pp. 2948-2956.

Ferlet-Cavrois, V.; Vizkelethy, G.; Paillet, P.; Torres, A.; Schwank, J. R.; Shaneyfelt, M. R., Baggio, J.; du Port de Pontcharra, J. & Tosti, L. (2004). Charge enhancement effect in NMOS bulk transistors induced by heavy ion irradiation—Comparison with SOI. *IEEE Transactions on Nuclear Science*, Vol. 51, No. 6, (Dec. 2004) pp. 3255–3262.

Ferlet-Cavrois, V.; Paillet, P.; McMorrow, D.; Torres, A.; Gaillardin, M.; Melinger, J. S.; Knudson, A. R.; Campbell, A. B.; Schwank, J. R.; Vizkelethy, G.; Shaneyfelt, M. R.; Hirose, K.; Faynot, O.; Jahan, C. & Tosti, L. (2005). Direct Measurement of Transient Pulses Induced by Laser Irradiation in Deca-Nanometer SOI Devices. *IEEE Transactions on Nuclear Science*, Vol. 52, No. 6, (Dec. 2005) pp. 2104-2113.

Fischetti, M. V. & Laux, S. E. (2001). Long-Range Coulomb Interactions in Small Si Devices. Part I: Performance and Reliability. *Journal of Applied Physics*, Vol. 89, No. 2, (2001) pp. 1205–1231.

Francis, P.; Colinge, J.P. & Beger, G. (1995). Temporal Analysis of SEU in SOI/GAA SRAMs. *IEEE Transactions on Nuclear Science*, Vol. 42, No.6, (Dec. 1995) pp. 2127-2137.

Frank, D.J.; Laux, S.E. & Fischetti., M.V. (1992). Monte Carlo simulation of a 30nm dual-gate MOSFET: How short can Si go?. *Proceedings of IEDM Technical Digest*, pp. 553-556, Washington, USA, Dec. 1992, IEEE.

Grubin, H.L.; Govindan, T.R.; Kreskovsky J.P. & Stroscio, M.A. (1993). Transport via the Liouville equation and moments of quantum distribution functions. *Solid-State Electronics*, Vol. 36, (Dec. 1993) pp. 1697–1709.

Guarini, K. W.; Solomon, P. M.; Zhang, Y.; Chan, K. K.; Jones, E. C.; Cohen, G. M.; Krasnoperova, A.; Ronay, M.; Dokumaci, O.; Bucchignano, J. J.; Cabral Jr., C.; Lavoie, C.; Ku, V.; Boyd, D. C.; Petrarca, K. S.; Babich, I. V.; Treichler, J.; Kozlowski, P. M.; Newbury, J. S.; D'Emic, C. P.; Sicina, R. M. & Wong, H. (2001). Triple-self-aligned, planar double-gate MOSFETs: Devices and circuits. *Proceedings of IEDM Technical Digest*, pp. 425-428, Washington, USA, Dec. 2001, IEEE.

Gusev, E. P.; Narayanan, V. & Frank, M. M. (2006). Advanced high-k dielectric stacks with polySi and metal gates: Recent progress and current challenges. *IBM Journal of Research and Development*, Vol. 50, No. 4/5, (2006) pp. 387–410.

Haensch, W.; Nowak, E. J.; Dennard, R. H.; Solomon, P. M.; Bryant, A.; Dokumaci, O.H.; Kumar, A.; Wang, X.; Johnson, J. B. & Fischetti, M. V. (2006). Silicon CMOS devices beyond scaling. *IBM Journal of Research and Development*, Vol. 50, No. 4/5, (2006) pp. 339-361.

Hamm, R. N.; Turner, J. E.; Wright, H. A. & Ritchie, R. H. (1979). Heavy ion track structure in Silicon. *IEEE Transactions on Nuclear Science*, Vol. 26, No. 6, (Dec. 1979) pp. 4892–4895.

Hansch, W.; Vogelsang, T.; Kirchner, R. & Orlowski, M. (1989). Carrier Transport Near the Si/SiO2 Interface of a MOSFET. *Solid State Electronics*, Vol. 32, No. 10, (Oct. 1989) pp. 839-849.

Hareland, S. A.; Jallepalli, S.; Shih, W.-K.; Wang, H.; Chindalore, G. L.; Tasch, A. F. & Maziar, C. M. (1998). A Physically-Based Model for Quantization Effects in Hole Inversion Layers. *IEEE Transactions on Electron Devices*, Vol. 45, No. 1, (Jan. 1998) pp. 179-186.

Harrison, S.; Coronel, P.; Leverd, F.; Cerutti, R.; Palla, R.; Delille, D.; Borel, S.; Descombes, S.; Lenoble, D.; Talbot, A.; Villaret, A.; Monfray, S.; Mazoyer, P.; Bustos, J.; Brut, H.; Cros, A.; Munteanu, D.; Autran, J-L. & Skotnicki, T. (2004). Highly performant double gate MOSFET realized with SON process. *Proceedings of IEDM Technical Digest*, pp. 449-452, Washington, USA, Dec. 2004, IEEE.

Hiramoto, T.; Saitoh M.; & Tsutsui, G. (2006). Emerging nanoscale Silicon devices taking advantage of nanostructure physics. *IBM Journal of Research and Development*, Vol. 50, No. 4/5, (2006) pp. 411-418.

Hisamoto, D.; Kaga, T.; Kawamoto, Y. & Takeda, E. (1989). A fully depleted lean-channel transist or (DELTA)-a novel vertical ultra thin SOI MOSFET. *Proceedings of IEDM Technical Digest*, pp. 833–836, Washington, USA, Dec. 1989, IEEE.

Houssa, M. (2004). *Fundamental and Technological Aspects of High-k Gate Dielectrics*, Institute of Physics, London.

ITRS 2009. International Technology Roadmap for Semiconductors. Available online: http://public.itrs.net.

Kedzierski, J.; Nowak, E.; Kanarsky, T.; Zhang, Y.; Boyd, D.; Carruthers, R.; Cabral, C.; Amos, R.; Lavoie, C.; Roy, R.; Newbury, J.; Sullivan, E.; Benedict, J.; Saunders, P.; Wong, K.; Canaperi, D.; Krishnan, M.; Lee, K.; Rainey, B. A.; Fried, D.; Cottrell, P.; Wong, H. P.; Ieong, M. & Haensch, W. (2002). Metal-gate FinFET and fully-depleted SOI devices using total gate silicidation. *Proceedings of IEDM Technical Digest*, pp. 247-250, Washington, USA, Dec. 2002, IEEE.

Kobetich, E. J. & Katz, R. (1968). Energy Deposition by Electron Beams and δ Rays. *Physical Review*, Vol. 170, No. 2, (1968) pp. 391-396.

Jiao, Z. & Salama, C. A. T. (2001). A fully depleted Δ-channel SOI nMOSFET. *Proceedings of the Electrochemical Society*, Vol. 2001–3, (2001) pp. 403–408.

Jiménez, D.; Iniguez, B.; Suné, J.; Marsal, L.F.; Pallarès, J.; Roig, J. & Flores, D. (2004). Continuous Analytic I–V Model for Surrounding-Gate MOSFETs. *IEEE Electron Device Letters*, Vol. 25, No. 8, (Aug. 2004) pp. 571-573.

Majkusiak, B.; Janik, T. & Walczak, J. (2002). Semiconductor Thickness Effects in the Double-Gate SOI MOSFET. *IEEE Transactions on Electron Devices*, Vol. 45, No. 5, (May 2002) pp. 1127-1134.

Martin, R. C.; Ghoniem, N. M.; Song, Y. & Cable, J. S. (1987). "The size effect of ion charge tracks on single event multiple-bit upset", *IEEE Transactions on Nuclear Science*, Vol. 34, No. 6, (1987) pp. 1305–1309.

Massengill, L. W.; Kerns, D. V.; Kerns, S. E. & Alles, M. L. (1990). Single-Event Charge Enhancement in SOI Devices, *IEEE Electron Device Letters*, Vol. EDL-11, (Feb. 1990) pp. 98-99.

Munteanu, D. & Autran, J.L. (2003). Two-dimensional Modeling of Quantum Ballistic Transport in Ultimate Double-Gate SOI Devices. *Solid State Electronics*, Vol. 47, No. 7, (2003) pp. 1219-1225.

Munteanu, D.; Ferlet-Cavrois, V.; Autran, J.L.; Paillet, P.; Baggio, J.; Faynot, O.; Jahan, C. & Tosti, L. (2006). Investigation of Quantum Effects in Ultra-Thin Body Single- and Double-Gate Devices Submitted to Heavy Ion Irradiation. *IEEE Transactions on Nuclear Science*, Vol. 53, No. 6, (Dec. 2006) pp. 3363-3371.

Munteanu, D.; Autran, J.L.; Ferlet-Cavrois, V.; Paillet, P.; Baggio, J.; & Castellani, K. (2007). 3-D Quantum Numerical Simulation of Single-Event Transients in Multiple-Gate Nanowire MOSFETs. *IEEE Transactions on Nuclear Science*, Vol. 54, No. 4, (Aug. 2007) pp. 994-1001.

Munteanu, D. & Autran, J.L. (2008). Modeling of digital devices and ICs submitted to transient irradiations. *IEEE Transactions on Nuclear Science*, Vol. 55, no. 4, (Aug. 2008) pp. 1854-1878.

Oldiges, P.; Dennard, R.; Heidel, D.; Klaasen, B.; Assaderaghi, R. & Ieong, M. (2000). Theoretical determination of the temporal and spatial structure of α-particle induced electron-hole pair generation in Silicon. *IEEE Transactions on Nuclear Science*, Vol. 47, No. 6, (Dec. 2000) pp. 2575–2579.

Park, J.T.; Colinge, J.P. & Diaz, C.H. (2001). Pi-Gate SOI MOSFET. *IEEE Electron Device Letters*, Vol. 22, No.8, (2001) pp.405-406.

Park, J.T. & Colinge, J.P. (2002). Multiple-gate SOI MOSFETs: device design guidelines. *IEEE Transactions on Electron Devices*, Vol. 49, No. 12, (Dec. 2002) pp. 2222-2229.

Rim, K.; Hoyt, J. L. & Gibbons, J. F. (1998). Transconductance Enhancement in Deep Submicron Strained-Si 12-MOSFETs. *Proceedings of IEDM Technical Digest*, pp. 707–710, Washington, USA, Dec. 1998, IEEE.

Roche, P. (1999). Etude du basculement induit par une particule ionisante dans une mémoire statique en technologie submicronique, Phd Thesis, 1999, in french.

Schwank, J. R.; Ferlet-Cavrois, V.; Shaneyfelt, M. R.; Paillet, P. & Dodd, P. E. (2003). Radiation effects in SOI technologies. *IEEE Transactions on Nuclear Science*, Vol. 50, No. 3, (Jun. 2003) pp. 522–538.

Stapor, W. J. & McDonald, P. T. (1988). Practical approach to ion track energy distribution. *Journal of Applied Physics*, Vol. 64, No. 9, (1988) pp. 4430-4434.

Stern, F. (1972). Self-consistent results for n-type Si inversion layers. *Physical Review B*, Vol. 5, (Jun. 1972) pp. 4891–4899.

Sentaurus (2009). Sentaurus TCAD Manuals, Synopsis, 2009.

Taur, Y.; Buchanan, D.; Chen, W.; Frank, D.; Ismail, K.; Lo, S.-H.; Sai-Halasz, G.; Viswanathan, R.; Wann, H.-J. C.; Wind, S. & Wong, H.-S. (1997). CMOS scaling into the nanometer regime. *Proceedings of IEEE*, Vol. 85, (1997) pp. 486–504.

Taur, Y. & Ning, TH. (1998). *Fundamentals of Modern VLSI Devices*. Cambridge Univ. Press, Cambridge, UK.

van Dort, M. J.; Woerlee, P. H. & Walker, A. J. (1994). A simple model for quantization effects in heavily-doped silicon MOSFET's at inversion conditions. *Solid-State Electronics*, Vol. 37, No. 3, (1994) pp. 411-414.

Wettstein, A.; Schenk, A. & Fichtner, W. (2002). Quantum device-simulation with Density-Gradient model. *IEEE Transactions on Electron Devices*, Vol. 48, No. 2, (Feb. 2002) pp. 279-284.

Yang, F.; Chen, H.; Chen, F.; Huang, C.; Chang, C.; Chiu, H.; Lee, C.; Chen, C.; Huang, H.; Chen, C.; Tao, H.; Yeo, Y.; Liang, M. & Hu, C. (2002). 25nm CMOS Omega FETs. *Proceedings of IEDM Technical Digest*, pp. 255-258, Washington, USA, Dec. 2002, IEEE.

4

Two-Fluxes and Reaction-Diffusion Computation of Initial and Transient Secondary Electron Emission Yield by a Finite Volume Method

Asdin Aoufi[1] and Gilles Damamme[2]
[1]*SMS/RMT, PECM, UMR CNRS 5146, Ecole des Mines de Saint-Etienne, 158 cours Fauriel, 42023 Saint-Etienne Cedex*
[2]*CEA -DAM/DIF, 91680 Bruyères le Châtel France*

1. Introduction

Dielectric breakdown in insulating materials is related to fast relaxation of trapped charges according to (G. Damamme & Reggi, 1997) and is of practical importance since it damages electronic devices (Levy, 2002). In fact, it is known that the secondary electron emission yield (*denoted by see*) is one of the key parameters for dielectric materials. Moreover, *see* is the driving parameter of electric charging which can lead to electric breakdown. To study this phenomena, the behaviour of an insulator submitted to an electron beam irradiation is considered. This has led to a significant number of experimental studies since the discovery of this phenomena. Although several modelling such as in (I.A.Glavatskikh & Fitting, 2001),(Fitting, 1974) and (H.-J. Fitting & Wild, 1977) are available in literature, they do not provide a simple method to compute the initial *see* yield from the penetration depth of the incident electron beam, and some material characteristics. The purpose of the work detailed in the book chapter is to describe such a simple modelling related to electron/matter interaction for low values of incident electron beam's energy and the tight coupling between modelling, numerical analysis and comparison with some experimental results.
This book chapter presents in a unified manner, published and new results. We focuss the presentation on our numerical/software approach, and comparison with experimental results.
In section 2, we propose a new modelling for the initial see yield computation. The main contribution is that we have reanalyzed from a mathematical point of view the modelling, stating that there exists a unique solution, which is uniformly bounded and which fulfill a maximum principle. From a numerical point of view, we have used a classical upwind finite-volume scheme, and shown the existence, uniqueness and discrete maximum principle for the discrete numerical solution. Finally a new asymptotic expression for the expression of see yield for large values of the electron beam energy is presented and discussed.
In section 3, we show that the computation of the initial see yield by a two-fluxes method, which requires the solution of a set of two coupled differential equations described in section 2, can in fact be reformulated into a single reaction-diffusion problem, which is much easier to solve from a computational point of view, since a single tri-diagonal matrix has to be inverted.

This appears to be a new result in this field.
In section 4, we present a model which computes the evolution of transient secondary electron emission yield. It is based upon a set of conservation laws which expresses the trapping of electrons and holes coupled with the electric field. From a numerical point of view, we apply a fully implicit scheme and uses a simple, fixed point technique to the solves the coupled set of discrete equations, and use a refined grid near the interface where the electron beam penetrates the sample. This enhances the quality of the numerical simulation and reduces significantly the elapsed computational time compared to Fitting's works (I.A.Glavatskikh & Fitting, 2001),(Fitting, 1974) and (H.-J. Fitting & Wild, 1977)¿Q, which is constrained by the fixed mesh spacing used in the presentation of his modelling. Moreover our numerical scheme uses the conservative finite-volume method, and we have proved formally that some discrete maximum principle occur which provides confidence in our numerical work. Some comparison between numerical computations and experimental work by G. Moya (IM2MP, Marseille, France) and K. Zarbout (IM2MP, Marseille, France, and LamaCop, Sfax, Tunisia) are presented.
In section 5, we extend the reformulation of the two-fluxes modelling presented in section 4 into a reaction-diffusion modelling. The main strength of this new approach is the ability to be extended easily in two spatial dimensions, while it is more difficult to extend in two-spatial dimensions the two-fluxes approach borrowed from the radiative transfer (**?**), hence this new approach seems more promising.
In section 6, we present the main architecture of our numerical software **sirena**.
We conclude this chapter with section 7, which summarizes the obtained results and draws some perspective of future work.

2. Initial see computation by a two-fluxes method

This section presents a modelling describing the generation of secondary electrons. The slowdown of primary electrons creates free electrons/holes pairs. A diffusion movement of these particles occurs. Some of the secondary electrons generated near the surface could be emitted when they are not trapped before.

2.1 Mathematical modelling

Let $C = \{e+,e-\}$, the charge transport of current fluxes $(j_c(z))_{c\in C}$ is described by a one-dimensional system of coupled linear differential equations along the z-axis. A two-fluxes method derived from radiative transfert theory is used which splits the electron current $j_e(z)$ into forward $j_{e+}(z)$ and backward $j_{e-}(z)$ contributions such that the algebraic currents verify $j_e(z) = j_{e+}(z) - j_{e-}(z)$. The coupling between $j_{e+}(z)$ and $j_{e-}(z)$ fluxes is written and takes into account diffusion.

$$-\frac{dj_{e-}(z)}{dz} + \left(\sigma_{e-}^{diff} + \sigma_e^{abs}\right) j_{e-}(z) = S_e(z) + \sigma_{e+}^{diff}\, j_{e+}(z), \tag{1}$$

with boundary condition $j_{e-}(L) = 0$, and

$$\frac{dj_{e+}(z)}{dz} + \left(\sigma_{e+}^{diff} + \sigma_e^{abs}\right) j_{e+}(z) = S_e(z) + \sigma_{e-}^{diff}\, j_{e-}(z). \tag{2}$$

with boundary condition $j_{e+}(0) = (1-\kappa)\ j_{e-}(0)$, where $\kappa \in [0,1]$ is the transmission coefficient, σ_e^{abs},$\sigma_{e\pm}^{diff}$ are respectively absorption and diffusion cross sections. This

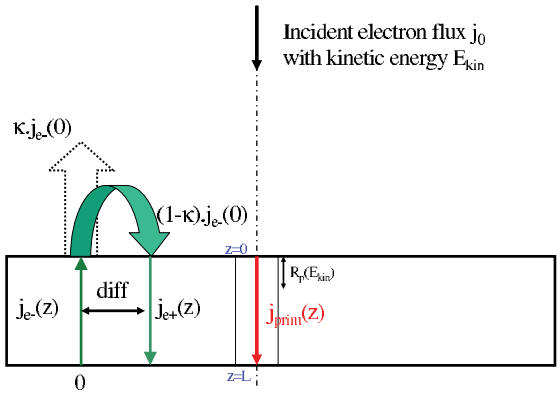


Fig. 1. Scheme of the modelling assuming that there is no backscattered electrons and where R_p is the penetration thickness and L is the dielectric thickness. The incident electron flux j_0 has a kinetic energy E_{kin}.

mathematical modelling is used to analyse the sensibility of the true secondary electron emission yield see^* defined by the expression

$$see^* = \kappa \frac{j_{e-}(0)}{j_0}. \tag{3}$$

with respect to the relative importance of charges absorption/diffusion inside the material, where j_0 is the current density of primary electrons (the backscattered electrons being excluded) and κ is the transmission coefficient.

2.2 Existence-uniqueness of the formal solution

The following proposition was presented in (Aoufi & Damamme, 2009)

Proposition 21 *Denoting the constant* $\sigma_c = \sigma_e^{abs} + \sigma_c^{diff}$, *and under the assumption that* $[z \mapsto S_e(z)]$ *is continuous over* Ω, *then*

– *the problem has a unique solution* $(j_{e-}(z), j_{e+}(z))$ *for* $z \in \Omega$ *given by the coupled system*

$$j_{e-}(z) = \int_z^L \left(S_e(E_{kin}, u; j_0) + \sigma_{e+}^{diff} j_{e+}(u) \right) e^{\sigma_{e-}(z-u)} du. \tag{4}$$

$$j_{e+}(z) = j_{e+}(0).e^{-\sigma_{e+}z} + \int_0^z \left(S_e(E_{kin}, u; j_0) + \sigma_{e-}^{diff} j_{e-}(u) \right) e^{\sigma_{e+}(u-z)} du. \tag{5}$$

– *Moreover there exits a constant* $C\left(\|S_e\|_{L^\infty(\Omega)}, E_{kin}, L, \sigma_e^{abs}, \sigma_c^{diff} \right) > 0$ *such that for* $z \in \Omega$

$$0 \le j_{e-,+}(z) \le j_0.C\left(\|S_e\|_{L^\infty(\Omega)}, E_{kin}, j_0, L, \sigma_e^{abs}, \sigma_c^{diff} \right)$$

2.3 Asymptotic expression for the secondary electron emission yield

Proposition 22 *We define the transfer cross section* $\sigma_e^{trans} = \sigma_e^{abs} + \sigma_{e+}^{diff} + \sigma_{e-}^{diff}$, *Under the asymption that* $1 \ll \sigma_e^{abs}.R_p(E_{kin})$ *then*

$$see^* \simeq \kappa \frac{S_e(z=0, E_{kin})}{\sigma_e^{abs}} \frac{\sigma^*}{\sigma_e^{trans}} \frac{\sigma^*}{(1-\frac{\kappa}{2})\sigma_e^{abs} + \frac{\kappa}{2}\sigma^*}. \tag{6}$$

where

$$\Delta\sigma_e^{diff} = \sigma_{e+}^{diff} - \sigma_{e-}^{diff},\ 2\sigma^* = \Delta\sigma_e^{diff} + \sqrt{\left(4\sigma_e^{abs}\sigma_e^{trans} + \left(\Delta\sigma_e^{diff}\right)^2\right)}. \tag{7}$$

Fig.(2) represents a comparison between asymptotic formula given by Eq.(6) and computation of *see* yield from Eq.(3). A good agreement is observed for high values of E_{kin}. Other computational results are given in (Aoufi & Damamme, n.d.).

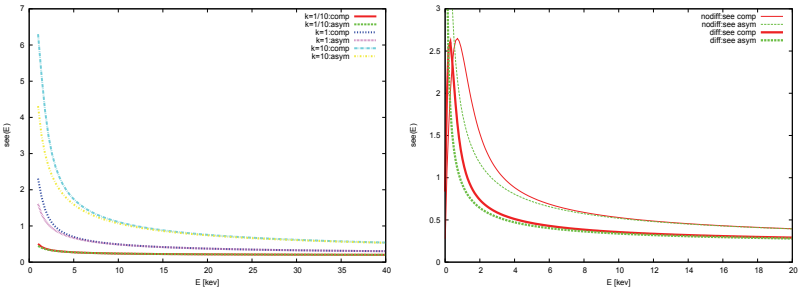


Fig. 2. (a) Comparison between computed and asymptotic expression for see yield in the case where $\sigma_{e+}^{diff} = k\sigma_e^{diff}$ and $\sigma_{e-}^{diff} = k^{-1}\sigma_e^{diff}$ with $k \in \{1/10, 1, 10\}$. (b) Comparison between numerical values and asymptotic expression of $see^*(E_{kin})$ as a function of E_{kin} [kev] in both cases with and without diffusion (for $\sigma_c^{diff} = \sigma_{\overline{c}}^{diff}$).

2.4 Numerical scheme

A vertex-centered conservative finite-volume discretization of the governing equation with an adequate upwind technique is defined. The domain Ω is decomposed into a set of I control volumes $\Omega_i = [z_i, z_{i+1}]$ with length $h_{i+\frac{1}{2}}$. The discrete unknown at grid point z_i related to j_c is denoted $j_c|_i$.

Proposition 23 *Denoting* $S_{i+\frac{1}{2}} = S_e\left(z_{i+\frac{1}{2}}\right)$, *where* $z_{i+\frac{1}{2}} = \frac{1}{2}(z_i + z_{i+1})$, *then*

– *the scheme obtained after integrating the forward linear equation over* $\Omega_{i+\frac{1}{2}}$ *is written for each cell index i,*

$$\begin{aligned} &\frac{j_{e+}|_{i+1} - j_{e+}|_i}{h_{i+\frac{1}{2}}} + \left(\sigma_e^{abs} + \sigma_{e+}^{diff} + \sigma_{e-}^{diff}\right) \cdot j_{e+}|_{i+1} \\ = \;& S_{i+\frac{1}{2}} + \sigma_{e-}^{diff}\left(j_{e-}|_{i+1} + j_{e+}|_i\right) \end{aligned} \tag{8}$$

– *using a similar computation, the scheme for the backward linear equation leads to the discrete equation*

$$\begin{aligned} &-\frac{j_{e-}|_i - j_{e-}|_{i-1}}{h_{i-\frac{1}{2}}} + \left(\sigma_e^{abs} + \sigma_{e+}^{diff} + \sigma_{e-}^{diff}\right) \cdot j_{e-}|_{i-1} \\ = \;& S_{i-\frac{1}{2}} + \sigma_{e+}^{diff}\left(j_{e+}|_{i-1} + j_{e-}|_i\right). \end{aligned} \tag{9}$$

– *The linear system with the discrete unknowns* $(j_{e-}|_i)_{1\le i\le I+1}, (j_{e+}|_i)_{1\le i\le I+1}$ *has a unique solution,*

– *which verify a discrete maximum principle for a suitable constant* $C > 0$

$$0 \leq j_{e-,+}|_i \leq C \tag{10}$$

2.5 Numerical simulations

Fig.(3) shows that in a suitable normalized representation, the evolution of secondary electron emission yield has a similar shape for different expressions of the penetration depth radius.

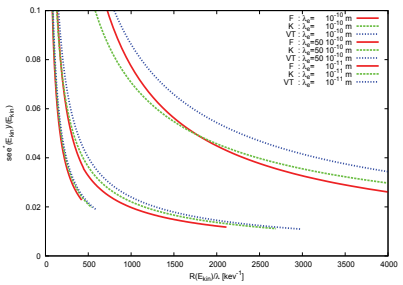


Fig. 3. Reduced variables evolution of $\frac{see^*(E_{kin})}{E_{kin}}$ in eV^{-1} as a function of $\frac{R(E_{kin})}{\lambda_e^{abs}}$ for the three penetration radius $R_p(E)$ laws, Vyatskin and Trunev (VT), Kanaya (K) and Fitting (F) for differents values of λ_e^{abs}.

3. Initial see computation by a reaction-diffusion method

3.1 Mathematical modelling

In order to reformulate Eq.(1)-(2) into a reaction diffusion equation, let us define the number of free electrons per unit volume $n_e(z)$ such that for $z \in [0,L]$, $v_e n_e(z) = j_{e+}(z) + j_{e-}(z)$ with v_e the mean absolute velocity of charge carriers and the overall transfert cross-section σ_e^{trans} according to $\sigma_e^{trans} = \sigma_e^{abs} + 2\sigma_e^{diff}$. After summation and substraction of Eq(1)-(2), and using the definition of $j_e(z)$ and $n_e(z)$ one obtain that

$$\frac{dj_e(z)}{dz} = 2S_e(z) - <\sigma_e^{abs} v_e> n_e(z),\ v_e \frac{dn_e(z)}{dz} = -j_e(z)\ \sigma_e^{trans}. \tag{11}$$

Defining the diffusion coefficient $D_e = \frac{v_e}{\sigma_e^{trans}}$ with respect to the transfer equation leads to the fact that the current flux $j_e(z)$ follows a Fick-type law : $j_e(z) = -D_e \frac{dn_e(z)}{dz}$. Pluging this expression into Eq.(3) leads to the reaction-diffusion equation with Fourier type boundary conditions

$$-D_e \frac{d^2 n_e(z)}{dz^2} = 2S_e(z) - <\sigma_e^{abs} v_e> n_e(z) \tag{12}$$

$$D_e \frac{dn_e(0)}{dz} = \frac{\kappa}{2-\kappa} n_e(0) v_e, \tag{13}$$

$$D_e \frac{dn_e}{dz}(L) = -n_e(L) v_e \tag{14}$$

The reformulation of Eq.(3) in terms of the number of trapped electrons $n_e(z)$ is easily obtained thanks to Eq.(11) and is such that :

$$see^* = \frac{\kappa}{2-\kappa}\frac{n_e(0)\,v_e}{j_0} \tag{15}$$

It is worth mentioning that in the case of electron transport, the velocity v_e can be obtained thanks to the equation $\frac{1}{2}m_e v_e^2 = E_e$ where the energy E_e was between $1\text{eV} - 3\text{eV}$, and is related to value of the gap energy.

3.2 Numerical scheme

A cell-centered finite-volume scheme on a geometrically refined grid near interface $z = 0$ is used. We prove the existence and uniqueness of the discrete solution that is computed by the inversion of a sparse tridiagonal matrix thanks to the classical Thomas algorithm -i.e. Gauss method for tridiagonal matrices-. We prove a discrete maximum principle, thanks to the M-matrix property of the tri-diagonal matrix.
We use a classical cell-centered finite volume approximation on computational domain Ω. A set of non-uniformly spaced grid points $(z_i)_{1\le i\le i+1}$ is given, and is such that $z_1 = 0 < \cdots < z_i < \ldots z_{I+1} = L$. We denote by h_i the length of control volume $\Omega_i = [z_i, z_{i+1}]$ and n_i the mean of $n(z)$ over control volume Ω_i. There are $I+1$ nodes, but I control volumes.

Proposition 3.1 *The finite-volume discretization of Eq.(14) leads to the following linear system*

$$\begin{pmatrix} b_1 & c_1 & & & \\ & \ddots & & & \\ & a_i & b_i & c_i & \\ & & & \ddots & \\ & & & a_I & b_I \end{pmatrix} \begin{pmatrix} n_1 \\ \vdots \\ n_i \\ \vdots \\ n_I \end{pmatrix} = \begin{pmatrix} 2h_1 S_1 \\ \vdots \\ 2h_i S_i \\ \vdots \\ 2h_I S_I \end{pmatrix}$$

with

$$a_1 = 0, \quad b_1 = \frac{D_e}{h_1 + \frac{h_2}{2}} + \frac{\kappa}{2-\kappa}v_e + < \sigma_e^{abs} v_e > h_1, \quad c_1 = -\frac{D_e}{h_1 + \frac{h_2}{2}}, \tag{16}$$

and for $i \in [1, I-1]$

$$a_i = \frac{D_e}{\frac{h_i + h_{i-1}}{2}}, \quad b_i = -\frac{D_e}{\frac{h_i + h_{i+1}}{2}} + \frac{D_e}{\frac{h_i + h_{i-1}}{2}} + < \sigma_e^{abs} v_e > h_i, \quad c_i = -\frac{D_e}{\frac{h_i + h_{i+1}}{2}}, \tag{17}$$

and

$$a_I = 0, \quad b_I = -\frac{D_e}{h_I + \frac{h_{I-1}}{2}}, \quad c_I = \frac{D_e}{h_I + \frac{h_{I-1}}{2}} + v_e. \tag{18}$$

which has a unique **positive solution.**

The discretization of Eq.(14) over control volume Ω_i leads to the following discrete equation which is specialized if $i = 1$, or $1 < i < I$ or $i = I$

$$-D_e\frac{n_2-n_1}{h_1+\frac{h}{2}}+\frac{\kappa}{2-\kappa}n_1 v_e = 2h_1 S_1+<\sigma_e^{abs}v_e>n_1 h_1 \quad (19)$$

$$-D_e\frac{n_{i+1}-n_i}{\frac{h_i+h_{i+1}}{2}}+D_e\frac{n_i-n_{i-1}}{\frac{h_i+h_{i-1}}{2}} = 2h_i S_i+<\sigma_e^{abs}v_e>n_i h_i \quad (20)$$

$$v_e n_I+D_e\frac{n_I-n_{I-1}}{h_I+\frac{h_{I-1}}{2}} = 2h_I S_I+<\sigma_e^{abs}v_e>n_I h_I \quad (21)$$

The matrix of the linear system that is used to compute $(n_i)_{1\leq i\leq I+1}$ is an M-matrix, since

- $b_i \geq 0,\ a_i, c_i \leq 0, b_i \geq |a_i|+|c_i|$,
- and $2h_i S_i \geq 0$.

we conclude that it is invertible and that the unique solution of the linear system is positive.

3.3 Numerical simulations

The evolution of $n_e(z)$ is depicted in Fig.(4). It is seen that $n_e(z) \geq 0$, and that its shape is closely influenced by the expression used for $S_e(z)$.

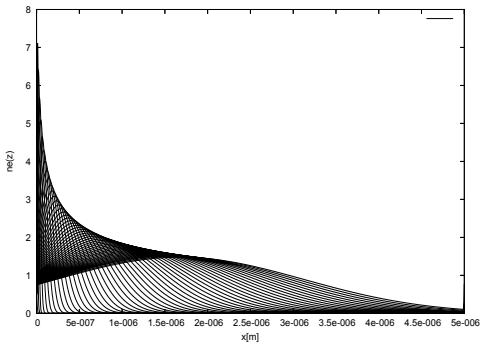


Fig. 4. Spatial distribution of $n_e(z)$ for various values E_{kin}.

4. Transient see computation by a two-fluxes method

In section 4, the modelling borrows from Fitting's papers, the absorption/diffusion cross-section expressions as a function of electric field. It differs mainly in the governing set of equations and in the numerical techniques that are used but also in the fact that some comparison between numerical computations and experimental work are done.

4.1 Mathematical modelling

The mathematical modelling expresses the coupling between electric field with electron/hole transport and describes the spatial and temporal charge trapping in an insulator submitted to an electron beam irradiation. The temporal evolution of the secondary electron emission is computed as a function of global trapped charge. It is given by a set of 7 nonlinear, coupled equations.

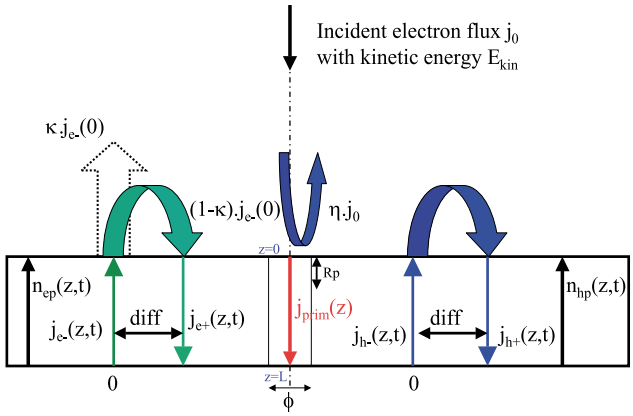


Fig. 5. Scheme of the modelling where R_p is the penetration depth, ϕ the diameter of the irradiated zone and L the dielectric thickness. The incident electron flux j_0 has a kinetic energy E_{kin}. Scheme of the modelling where $R_p(E_{kin})$ is the penetration depth, ϕ the diameter of the irradiated zone and L the dielectric thickness. The incident electron flux j_0 has a kinetic energy E_{kin}. The electron fluxes j_{e-}/j_{e+} and the hole fluxes j_{h-}/j_{h+} are coupled by diffusion.

4.1.1 Purpose of the modelling

The purpose of this modelling is to analyze the evolution of the global trapped charge, per unit surface, at time t, $Q_p(t)$ which is defined by:

$$Q_p(t) = \int_0^L \rho(z,t)\,dz + |e|\,n_s(t) \tag{22}$$

and the true secondary electron emission yield $see^*(t)$ expressed as:

$$see^*(t) = \kappa_s\left(n_s(t)\right)\frac{j_{e-}(0,t)}{j_0} \tag{23}$$

where the expression of $\kappa\left(n_s(t)\right)$ is given in the next subsubsection.
The proposed modelling describes the interaction between the number of trapped electrons $n_{ep}(z,t)$ and holes $n_{hp}(z,t)$ with the four current fluxes $(j_c)_{c\in\mathcal{C}}$, the current $j_{prim}(z)$ and the electric field $E(z,t)$.

4.1.2 Governing equation for the saturation effect of the surface trapping sites $n_s(t)$

Defining by N_s the number of trapping surface sites located at the interface $z=0$ and by σ_s their elementary cross section then the evolution of the number of traps per surface unit at time t, $n_s(t)$ follows the equation

$$|e|\,\frac{dn_s(t)}{dt} = j_{e-}(0,t)\left(\kappa - \kappa_s\left(n_s(t)\right)\right) \tag{24}$$

where $\kappa_s\left(ns(t)\right)$ is defined by:

$$\kappa(n_s(t)) = \kappa\exp\left(-\sigma_s\left(N_s - n_s(t)\right)\right) \tag{25}$$

Its contribution is especially important during the initial charge injection phase for an amount of time driven by the product $\sigma_s N_s$. The initial condition states that there are no surface trapped charges.

4.1.3 Governing equation for the electric field $\mathbf{E}(\mathbf{z},\mathbf{t})$

Here we assume that the electric field $E(z,t)$ depends only on the total charge density $\rho(z,t)$ which is different from the trapped charge density. Using a two-fluxes method it was assumed that only the trapped charge density contributed to the electric field. The local Maxwell-Gauss equation writes

$$\nabla.E\,(z,t) = \frac{\rho\,(z,t)}{\epsilon_0\epsilon_r}, \tag{26}$$

An electrostatic analysis taking into account polarization charges on the interface leads to

$$E\,(0,t) = -\frac{1}{\epsilon_0\epsilon_r\,(1+\varepsilon_r)}\;\frac{1}{1+\frac{\pi\phi^2/4}{L^2(1+\varepsilon_r)}}\;Q_p(t) - |e|\frac{n_s(t)}{\epsilon_0\epsilon_r}. \tag{27}$$

where the second factor is introduced as corrections due to image charges in the sample holder.

4.1.4 Governing equation for current fluxes $j_c\,(z,t)$

The following balance is written

$$d_c\frac{\partial j_c(z,t)}{\partial z} + \sum_{i=1}^{3} W_{c,i} = S_c(z) + \sigma_c^{diff}\,(E(z,t))\,j_c(z,t) \tag{28}$$

where

(a) $d_c\frac{\partial j_c(z,t)}{\partial z}$ is the gradient of forward/backward electron/hole flux,

(b) $W_{c,1} = \sigma_c^{diff}\,(E(z,t))\,j_c(z,t)$ is the flux loss by diffusion,

(c) $W_{c,2} = \left(\sigma_{pc}\left(N_{pc} - n_{cp}(z,t)\right)\right).j_c(z,t)$, is the flux loss by trapping on unoccupied trapping sites $N_{pc} - n_{cp}(z,t)$,

(d) $W_{c,2} = \sigma_{ac}n_{\hat{c}p}(z,t)j_c(z,t)$ is the flux loss by annihilation with trapped charge $\hat{c}$,

(e) $S_c(z)$ is a source term for the creation of electrons or holes induced by the slowdown of primary electrons.

(f) $\sigma_c^{diff}\,(E(z,t))\,j_c(z,t)$ is a positive source transport term by diffusion for dual charge of c, i.e. travelling in the opposite direction.

The boundary conditions are s_c sign dependant.

- For backward fluxes $j_{e-}\,(z,t)$ and $j_{h-}\,(z,t)$, the boundary condition at $z = L$ means that the bottom of the material has no charge injection: $\forall c \in \{e-,h-\}\;:\; j_c\,(L,t) = 0$.
- For forward fluxes, $j_{e+}\,(z,t)$ and $j_{h+}\,(z,t)$, the boundary condition at $z = 0$ means the continuity of the hole flux $j_{h+}\,(0,t) = j_{h-}\,(0,t)$, but a discontinuity for the electron flux $j_{e+}\,(0,t) = (1-\kappa)\,.j_{e-}\,(0,t)$.

4.1.5 Governing equation for the charge density $\rho\,(z,t)$

The temporal variation of the trapped charge density $\rho(z,t)$ follows the conservation law

$$\frac{\partial\rho\,(z,t)}{\partial t} + \nabla.j\,(z,t) = 0, \tag{29}$$

where the overall current flux of charges $j\,(z,t)$ is such that

$$j\,(z,t) \quad = \quad j_{e-}\,(z,t) - j_{e+}\,(z,t) - j_{h-}\,(z,t) + j_{h+}\,(z,t) - j_0\,j_{prim}\,(z)\,. \tag{30}$$

4.1.6 Governing equation for trapped charge $Q_p(t)$

The temporal evolution of total trapped charge $Q_p(t)$ is defined from the total charge density

$$\rho(z,t) = |e| \left(n_{ep}(z,t) - n_{hp}(z,t) \right) \tag{31}$$

thanks to the equation

$$\frac{dQ_p(t)}{dt} = \int_0^L \frac{d\rho(z,t)}{dt} = -j_{e+}(L,t) + j_{h+}(L,t) + j_0(1 - see(t)). \tag{32}$$

4.1.7 Governing equation for the trapped electrons $n_{ep}(z,t)$

The evolution of the number of trapped electrons $n_{ep}(z,t)$ follows the differential equation

$$\begin{aligned} |e| \frac{\partial n_{ep}}{\partial t}(z,t) &= \sigma_{pe} \left(N_{pe} - n_{ep}(z,t) \right) \left(j_{e+}(z,t) + j_{e-}(z,t) \right) \\ &\quad -\sigma_{ah} n_{ep}(z,t) \left(j_{h+}(z,t) + j_{h-}(z,t) \right) \end{aligned} \tag{33}$$

which expresses the balance between

– the number of electrons that are trapped, where σ_{pe} is the trapping cross section of the electrons and N_{pe} is the total number of electrons trapping sites,

$$\sigma_{pe} \left(N_{pe} - n_{ep}(z,t) \right) \left(j_{e+}(z,t) + j_{e-}(z,t) \right) \tag{34}$$

– and the number of trapped electrons present in the traps that are annihilated by free holes, where σ_{ah} is the annihilation cross section between trapped electrons and free holes.

$$\sigma_{ah} n_{ep}(z,t) \left(j_{h+}(z,t) + j_{h-}(z,t) \right). \tag{35}$$

The absolute value, $|e|$ of the electron charge in coulomb is introduced to transform the fluxes $(j_c(z,t))_{c \in C}$ expressed in Coulomb into fluxes expressed in carriers numbers.
The initial condition $n_{ep}(z,0) = 0$, means that there are no trapped electrons at the beginning of charge injection.

Remark 4.1 *In order to simplify the notations, we define the total flux* $j^T_{h,e}(z,t) = j_{h+/e+}(z,t) + j_{h-/e-}(z,t)$*, which is always positive, while the algebraic flux* $j_{h,e}(z,t) = j_{h+/e+}(z,t) - j_{h-/e-}(z,t)$ *can be negative. We introduce functions* $a(z,t)$ *and* $b(z,t)$ *that are defined by the expressions*

$$\begin{aligned} a(z,t) &= \left(\sigma_{pe} j^T_e(z,t) + \sigma_{ah} j^T_h(z,t) \right) / |e| \geq 0, \\ b(z,t) &= \left(\sigma_{pe} N_{pe} j^T_e(z,t) \right) / |e| \geq 0. \end{aligned}$$

then Eq.(33) can be rewritten

$$\frac{\partial n_{ep}}{\partial t}(z,t) + a(z,t) n_{ep}(z,t) = b(z,t) \tag{36}$$

A straightforward computation shows that the formal solution of Eq.(33) is given by

$$n_{ep}(z,t) = \int_0^t \exp^{\int_t^v a(z,u)\,du} b(z,v)\,dv \tag{37}$$

from which we can infer that $n_{ep}(z,t) \geq 0$.

4.1.8 Governing equation for the trapped holes $n_{hp}(z,t)$

The evolution of the number of trapped holes $n_{hp}(z,t)$ follows the differential equation

$$\begin{aligned} |e|\frac{\partial n_{hp}}{\partial t}(z,t) &= \sigma_{ph}\left(N_{ph}-n_{hp}(z,t)\right)\left(j_{h+}(z,t)+j_{h-}(z,t)\right) \\ &\quad -\sigma_{ae}n_{hp}(z,t)\left(j_{e+}(z,t)+j_{e-}(z,t)\right) \end{aligned} \tag{38}$$

which expresses the balance between

– the number of holes that are trapped, where σ_{ph} is the trapping cross section of the holes and N_{ph} is the total number of holes trapping sites,

$$\sigma_{ph}\left(N_{ph}-n_{hp}(z,t)\right)\left(j_{h+}(z,t)+j_{h-}(z,t)\right) \tag{39}$$

– and the number of trapped holes present in the traps that are annihilated by free electrons, where σ_{ae} is the annihilation cross section between trapped holes and free electrons.

$$\sigma_{ae}n_{hp}(z,t)\left(j_{e+}(z,t)+j_{e-}(z,t)\right) \tag{40}$$

The initial condition $n_{hp}(z,0)=0$, means that there are no trapped holes at the beginning of charge injection.
It is worth mentioning that trapped holes and trapped electrons have the same type of behaviour, so the governing equations are symmetrical, when one exchanges the index h with the index e.

4.2 Numerical scheme

The mathematical modelling expresses the nonlinear coupling between a set of seven equations with seven unknowns $E(z,t)$, $(j_c(z,t))_{c\in\mathcal{C}}$, $n_{ep}(z,t)$ and $n_{hp}(z,t)$. A straighforward computation leads to a formal expression of each unknown as a function of the others which involves spatial/temporal integrals and stiff exponential functions, but the non-linear coupling remain. We are therefore led to use a numerical discretization scheme to compute the solution of this one-dimensional nonlinear initial boundary value problem, expressed in conservation form.
We present a full implicit conservative finite volume scheme on a non uniform **staggered grid** used for the discretization of the governing set of equations on a geometrically refined grid near the interface z=0. The computational domain is Ω.

– Unknowns that are located at the **center of cell** Ω_i are $n_{ep}|_i^k$, $n_{hp}\Big|_i^k$, $\rho|_i^k$,

– Unknowns that are located at the **edges of cell** Ω_i **are** $E|_{i,i+1}^k$, $j_{e+}|_{i,i+1}^k$, $j_{e-}|_{i,i+1}^k$, $j_{h+}|_{i,i+1}^k$, $j_{h-}|_{i,i+1}^k$.

We use a backward Euler scheme, with constant time-step Δt, first order accurate in time, for the temporal discretization, and note $t_k = k.\Delta t$.

4.2.1 Discretization of the surface trapping sites $n_s(t)$ equation

A straightforward computation leads to the discrete equation

$$\frac{n_s^{k+1}-n_s^k}{\Delta t}=\frac{j_{e-}|_1^{k+1}}{|e|}\left(\kappa-\kappa_s(n_s^{k+1})\right). \tag{41}$$

The stiffness induced by the exponential term present in $\kappa_s(n_s^{k+1})$ requires a first order linearization, hence the iterative solution by a fixed point technique is given by

$$\frac{n_s^{k+1,p+1} - n_s^{k,p}}{\Delta t} = \frac{j_{e-}|_1^{k+1,p}}{|e|}(\kappa - w^p)), \tag{42}$$

$$w^p = \kappa_s(n_s^{k+1,p}) + \left(n_s^{k+1,p+1} - n_s^{k+1,p}\right)\kappa_s'(n_s^{k+1,p}). \tag{43}$$

Then the value of $n_s^{k+1,p+1}$ is easily determined.

4.2.2 Discretization of the charge density equation

Integrating Eq.(29) over control volume Ω_i leads to the discrete equation

$$h_i\left(\rho_i^{k+1} - \rho_i^k\right) + \Delta t\left(j_{i+1}^{k+1} - j_i^{k+1}\right) = 0. \tag{44}$$

A decoupled iterative solution thanks to the fixed point method leads to the computation, where p is the nonlinear iteration index

$$\rho_i^{k+1,p+1} = \rho_i^k - \frac{\Delta t}{h_i}\left(j_{i+1}^{k+1,p} - j_i^{k+1,p}\right). \tag{45}$$

4.2.3 Discretization of the trapped charge equation

The discretization is straightforward and leads to the equation

$$\frac{Q_p\big|^{k+1} - Q_p\big|^k}{\Delta t} = -\,j_{e+}|_{I+1}^{k+1} + j_{h+}|_{I+1}^{k+1} + j_0\left(1 - see^{k+1}\right). \tag{46}$$

A decoupled iterative solution thanks to the fixed point method leads to the computation, where p is the nonlinear iteration index

$$Q_p\big|^{k+1,p+1} = Q_p\big|^{k,p} + \Delta t\left(-\,j_{e+}|_{I+1}^{k+1,p} + j_{h+}|_{I+1}^{k+1,p} + j_0\left(1 - see^{k+1,p}\right)\right). \tag{47}$$

4.2.4 Discretization of the electric field equation

The discretization is straightforward and leads to the equation

$$E_{i+1}^{k+1} - E_i^{k+1} = \frac{h_i\rho_i^{k+1}}{\epsilon_0\epsilon_r}, \tag{48}$$

$$E_1^{k+1} = -\frac{1}{\epsilon_0\epsilon_r\left(1+\varepsilon_r\right)}\,\frac{1}{1+\frac{\pi\phi^2/4}{L^2(1+\varepsilon_r)}}\,Q_p^{k+1} - |e|\frac{n_s^{k+1}}{\epsilon_0\epsilon_r}. \tag{49}$$

A decoupled iterative solution thanks to the fixed point method leads to the computation, where p is the nonlinear iteration index.

$$E_{i+1}^{k+1,p+1} = E_i^{k+1} + \frac{h_i\rho_i^{k+1,p}}{\epsilon_0\epsilon_r}, \tag{50}$$

$$E_1^{k+1,p+1} = -\frac{1}{\epsilon_0\epsilon_r\left(1+\varepsilon_r\right)}\,\frac{1}{1+\frac{\pi\phi^2/4}{L^2(1+\varepsilon_r)}}\,Q_p^{k+1,p} - |e|\frac{n_s^{k+1,p}}{\epsilon_0\epsilon_r}. \tag{51}$$

4.2.5 Discretization of the trapped holes's number equation

In order to simplify the notations all the cross section terms are divided by the factor $|e|$.

Proposition 4.1 *The finite volume discretization of Eq.(38) over control volume Ω_i leads to the discrete equation*

$$\frac{n_{hp}\Big|_i^{k+1} - n_{hp}\Big|_i^{k}}{\Delta t} = \sigma_{ph}\left(N_{ph} - n_{hp}\Big|_i^{k+1}\right) j_h^T\Big|_i^{k+1} - \sigma_{ae}.\, n_{hp}\Big|_i^{k+1}.\, j_e^T\Big|_i^{k+1}. \tag{52}$$

which has a unique solution given by

$$n_{hp}\Big|_i^{k+1} = \frac{n_{hp}\Big|_i^{k} + \Delta t.\sigma_{ph}.N_{ph}.\, j_h^T\big|_i^{k+1}}{1 + \Delta t.\sigma_{ph}\, j_h^T\big|_i^{k+1} + \Delta t.\sigma_{ae}\, j_e^T\big|_i^{k+1}} \tag{53}$$

for which the discrete maximum principle holds

$$0 \leq n_{hp}\Big|_i^{k} \leq N_{ph}. \tag{54}$$

Moreover thanks to the fixed point technique, we have the following nonlinear iteration

$$n_{hp}\Big|_i^{k+1,p+1} = \frac{n_{hp}\Big|_i^{k} + \Delta t.\sigma_{ph}.N_{ph}.\, j_h^T\big|_i^{k+1,p}}{1 + \Delta t.\sigma_{ph}\, j_h^T\big|_i^{k+1,p} + \Delta t.\sigma_{ae}\, j_e^T\big|_i^{k+1,p}} \tag{55}$$

Let us construct the finite volume discretization. We integrate Eq.(**??**) over control volume $\Omega_i \times [t_k, t_{k+1}]$ to obtain

$$\begin{aligned} &\int_{\Omega_i}\left(n_{hp}(z,t^{k+1}) - n_{hp}(z,t^k)\right) dz \\ = \quad &\int_{t^k}^{t^{k+1}}\left(\int_{\Omega_i}\sigma_{ph}\left(N_{ph} - n_{hp}(z,t)\right).j_h^T(z,t) - \sigma_{ae}.n_{hp}(z,t).j_e^T(z,t)\right) dt \end{aligned} \tag{56}$$

A cell-centered approximation is used for n_{hp}, then

$$\int_{\Omega_i}\left(n_{hp}(z)^{k+1} - n_{hp}(z)^k\right) dz = h_i\left(n_{hp}\Big|_i^{k+1} - n_{hp}\Big|_i^{k}\right). \tag{57}$$

But a vertex-centered approximation is used for $j_{e\pm,h\pm}(z,t)$, so we apply a first order approximation to the integral, i.e. we evaluate the integrand,which is a fonction of $j_{e\pm,h\pm}(z,t)$ at $z = z_i$ while $n_{hp}(z,t)$ is a constant over $\Omega_i \times [t^k, t^{k+1}]$ and equal to $n_{hp}\Big|_i^{k+1}$, to obtain

$$\begin{aligned} &\int_{t^k}^{t^{k+1}}\left(\int_{\Omega_i}\sigma_{ph}\left(N_{ph} - n_{hp}(z,t)\right).j_h^T(z,t) - \sigma_{ae}.n_{hp}(z,t).j_e^T(z,t)\right) dt \\ = \quad &h_i\Delta t\left(\sigma_{ph}\left(N_{ph} - n_{hp}\Big|_i^{k+1}\right) j_h^T\Big|_i^{k+1} - \sigma_{ae}.\, n_{hp}\Big|_i^{k+1}.\, j_e^T\Big|_i^{k+1}\right). \end{aligned} \tag{58}$$

We now prove by induction that the discrete maximum principle holds.

– For $k = 0$, thanks to the initial condition, we have $\forall i \in [1, I],\ n_{hp}\Big|_i^0 = 0 \in \Big[0, N_{ph}\Big]$, so the condition is fulfilled.

– For a given time index k, let us assume that $n_{hp}\Big|_i^k \in \Big[0, N_{ph}\Big]$. Computation of $n_{hp}\Big|_i^{k+1}$ is given by expression

$$n_{hp}\Big|_i^{k+1} = \frac{n_{hp}\Big|_i^k + \Delta t.\sigma_{ph}\ N_{ph}.\ j_h^T\big|_i^{k+1}}{1 + \Delta t.\sigma_{ph}.\ j_h^T\big|_i^{k+1} + \Delta t.\sigma_{ae}.\ j_e^T\big|_i^{k+1}} = \alpha.\ n_{hp}\Big|_i^k + \beta.N_{ph} \tag{59}$$

where

$$\left(1 + \Delta t.\sigma_{ph}.\ j_h^T\Big|_i^{k+1} + \Delta t.\sigma_{ae}.\ j_e^T\Big|_i^{k+1}\right)\alpha \quad = \quad 1, \tag{60}$$

$$\left(1 + \Delta t.\sigma_{ph}.\ j_h^T\Big|_i^{k+1} + \Delta t.\sigma_{ae}.\ j_e^T\Big|_i^{k+1}\right)\beta \quad = \quad \Delta t.\sigma_{ph}.\ j_h^T\Big|_i^{k+1}. \tag{61}$$

but $\alpha \geq 0,\ \beta \geq 0,\ \alpha + \beta \leq 1$, hence $0 \leq n_{hp}\Big|_i^{k+1} \leq \max\left(n_{hp}\Big|_i^k, N_{ph}\right) = N_{ph}$. So the discrete maximum principle is verified for $n_{hp}\Big|_i^{k+1}$.

A similar result can be stated and proved for the discrete approximation of trapped electron's equation.

4.3 Numerical simulations

4.3.1 Analysis of the influence of E_{kin}

In this subsection we investigate the sensibility of $Q_p(t)$ and $ees(t)$ with respect to the energy of the primary electrons. We assume that N_{pe}, N_{ph} and cross sections $\sigma_{ae,h}$ are fixed.

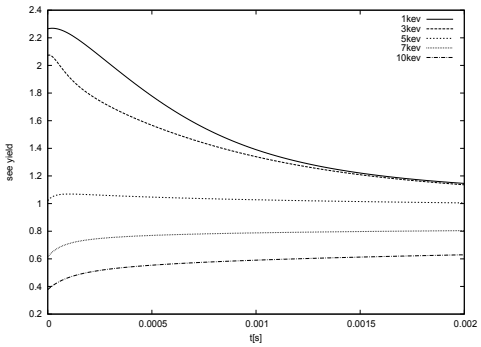


Fig. 6. Evolution of the secondary electron emission ratio as a function of time t, for various values of E_{kin}.

The behaviour of *see* with respect to E_{kin} is presented in Fig.6 and shows that for small values of E_{kin} the ratio starts from a value greater than one and decreases down to one very quickly. On the other hand for values of E_{kin} greater than 5 kev, the ratio starts below one, and strictly increases to the asymptotic value of one.

Spatial profiles of electron current $j_{e-,+}(z,t)$ and holes currents $j_{h-,+}(z,t)$ are similar in shape and amplitude, hence we only repesented in Fig.7. j_{e-} profiles. It is worth mentioning that

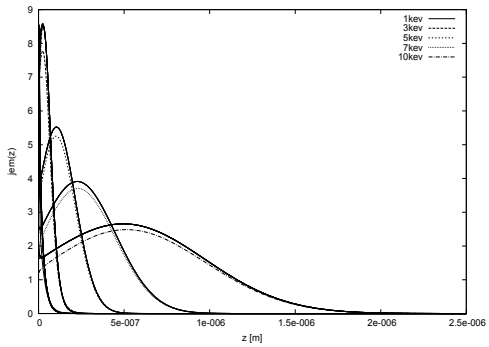


Fig. 7. Spatial distribution of electron current $j_{e-}(z,t)$ over domain $[0,L]$, for various values of E_{kin} at times $k \times 200.10^{-6}$, with $0 \leq k \leq 10$.

increasing E_{kin} induces that the maximum of $j_c(z,t)$ decreases and the profiles are diffused pushed towards $z = L$. This result is correlated with the shape of the source term $S_e(z)$ which varies accordingly when E_{kin} increases.
Trapped holes $n_{hp}(z,t)$ and trapped electrons $n_{ep}(z,t)$ are respectively represented in Fig.9 and Fig.8. The maximum of $n_{ep}(z,t)$ is smaller than the maximum of $n_{hp}(z,t)$ for each value of E_{kin} and always decreases when E_{kin} increases while the spatial profiles are smeared out.
The variation of $\log(ees(Qp))$ is represented in Fig.10 and shows a significant difference depending on the value of E_{kin}. When E_{kin} is below 5 kev, a strictly superlinear smooth decreasing profile is observed. On the other hand when E_{kin} is higher than 5 kev, the profiles have a high curvature.
Fig.11 is an enlarged representation of electric field $E(z,t)$. Three different domains can be observed for each value of E_{kin} and for each time step. A) a thin boundary layer located near the interface $z = 0$ where the $E(z,t)$ is stiff, B) a central part where $E(z,t)$ is oscillating and a third part near the interface $z = L$ where $E(z,t)$ is flat.

4.3.2 Analysis of the influence of the value of N_{pe} and N_{ph}

In this subsection we assume that the kinetic energy of the incident electron beam is constant and given the value of 4 kev, and study the influence of N_{pe}, and N_{ph} on the variation of $sse(t)$ and $Q_p(t)$. The trend observed in the numerical simulations appears independent of E_{kin}.

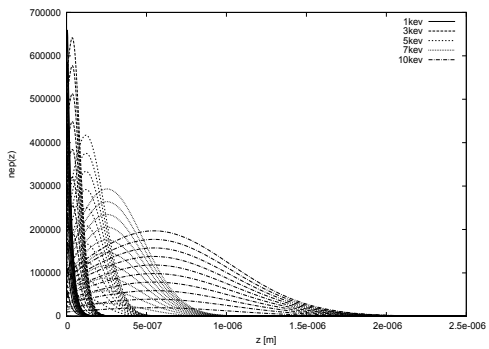


Fig. 8. Spatial distribution of trapped electrons $n_{ep}(z,t)$ over domain $[0,L]$, for various values of E_{kin} at times $k \times 200.10^{-6}$, with $0 \leq k \leq 10$.

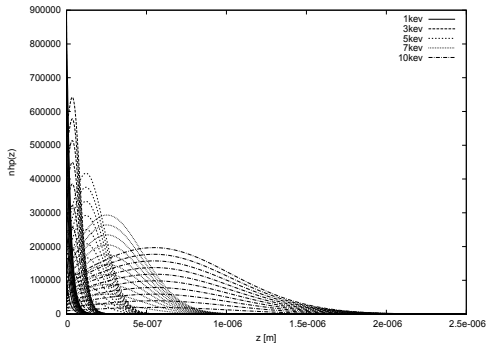


Fig. 9. Spatial distribution of trapped holes $n_{hp}(z,t)$ over domain $[0,L]$, for various values of E_{kin} at times $k \times 200.10^{-6}$, with $0 \leq k \leq 10$.

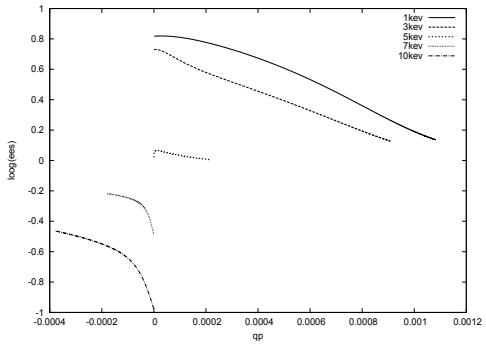


Fig. 10. Evolution of $\log(ees)$ as function Q_p for various values of E_{kin}.

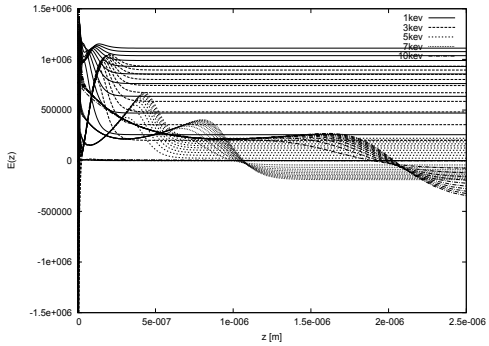


Fig. 11. Spatial distribution of electric field $E(z,t)$ over domain $[0,L]$, for various values of E_{kin} at times $k \times 200.10^{-6}$, with $0 \leq k \leq 10$.

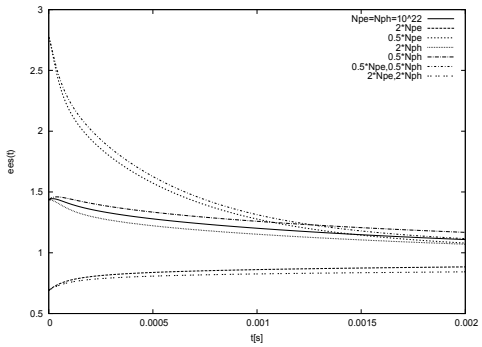


Fig. 12. Sensibility of the temporal evolution of see(t) when both values of N_{pe} and N_{ph} are varied simultaneously or independently for E=4keV.

As seen on Fig.12, reducing in/dependently N_{pe} and N_{ph} leads to a significant increase of the secondary electron emission ratio well above 1. On the other hand reducing in/dependently N_{pe} and N_{ph} leads to a significant decrease of the total trapped charge $Q_p(t)$.
The analysis of the evolution of $\log(ees(Q_p(t)))$ represented in Fig.13 shows that setting N_{pe} and varying N_{ph} induces that the curves are parallel, but converging to the same point where $Q_p = 0$. On the other hand, setting N_{ph} and varying N_{pe} above or below the initial value of N_{pe} leads to a much wider variation of *see* above or below 1 as seen on Fig.12.
As a conclusion, the most important parameter in these simulations appears to be N_{pe}.

4.3.3 Comparison between numerical computations and experimental results

Preliminary promising results are presented in Fig.14.

5. Transient see computation by a reaction-diffusion method

In section 5, we extend the reformulation of the two-fluxes modelling presented in section 4 into a reaction-diffusion modelling. The main strength of this new approach is the ability to be extended in two/three spatial dimensions. Moreover, it is more difficult to extend in two/three-spatial dimensions the two-fluxes approach borrowed from the radiative transfer (Chandrasekhar 1961), hence this new approach seems more promising.

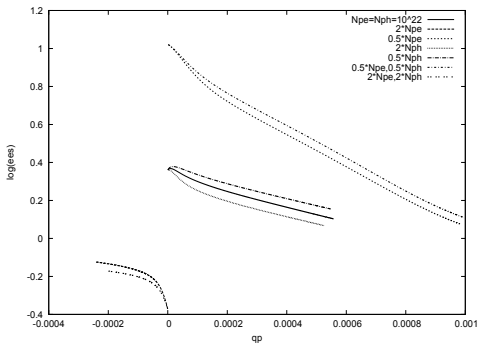


Fig. 13. Evolution of $\log(ees)$ as function Q_p when both values of N_{pe} and N_{ph} are varied simultaneously or independently for E=4keV.

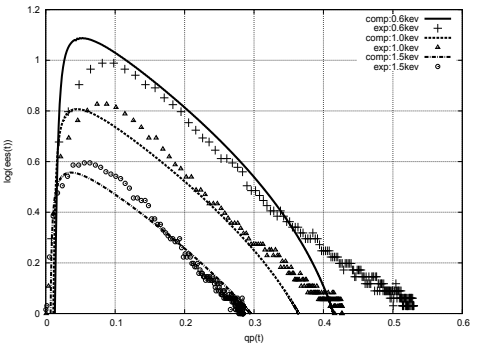


Fig. 14. Comparison between experimental and computed values of secondary electron emission yield $see(t)$ versus trapped charge $Q_p(t)$.

5.1 Mathematical Modelling

We present the modelling composed of a set of two, one dimensional reaction-diffusion equations for electrons/holes coupled with Gauss equation for the electric field and an equation for trapped electrons/holes evolution.

5.1.1 Governing equation for the electric field $E(z,t)$

The local conservation equation for the electric field $E(z,t)$ writes

$$\nabla.E\left(z,t\right)=\frac{\rho\left(z,t\right)}{\varepsilon_0\varepsilon_r} \tag{62}$$

The boundary condition at $z=0$, derived from an electrostatic analysis is written

$$E\left(0,t\right)=-\frac{1}{\varepsilon_0\varepsilon_r\left(1+\varepsilon_r\right)}\int_0^L\rho\left(z,t\right)dz \tag{63}$$

5.1.2 Governing equation for charge density $\rho(z,t)$

We define the overall current flux $j_T\left(z,t\right)=j_e\left(z,t\right)-j_h\left(z,t\right)-j_0\,j_{prim}\left(z\right)$. The conservation law expressing the evolution of charge density $\rho(z,t)$ is written

$$\frac{\partial\rho\left(z,t\right)}{\partial t}+\nabla.j_T\left(z,t\right)=0 \tag{64}$$

with initial condition $\rho\left(z,0\right)=0$ expressing the lack of charge.

5.1.3 Governing equations for the number of free charges $\left(n_c\left(z,t\right)\right)_{c\in C}$ and current fluxes $j_c(z,t)$

For charge c, the number of trapped charge, either electrons or holes, $n_c\left(z,t\right)$ is based on the following balance equation

$$\frac{\partial n_c}{\partial t}\left(z,t\right)+\nabla.j_c\left(z,t\right)=2S_c\left(z\right)-\left(\sigma_c^{abs}v\right)n_c\left(z,t\right) \tag{65}$$

The current flux for charge c $j_c\left(z,t\right)$ is related to the number of free charges $n_c\left(z,t\right)$ thanks to the equation

$$j_c\left(z,t\right)=-D_c\nabla n_c\left(z,t\right)+n_c\left(z,t\right)\ \mu_c s_c E\left(z,t\right). \tag{66}$$

For a given charge $c \in C$, its sign s_c is set to be $+1$ for the holes and -1 for the electrons and its mobility is represented by $\mu_c \geq 0$. The diffusion coefficient D_c is defined by $D_c = \frac{v_c}{\sigma_c^{trans}}$, where $\sigma_c^{trans} = \sigma_c^{abs} + 2\sigma_c^{diff}$. Hence the partial differential equation related to the evolution of $n_c(z,t)$ is rewritten

$$\frac{\partial n_c}{\partial t}(z,t) + \nabla.(-D_c \nabla n_c(z,t) + n_c(z,t)\ \mu_c s_c E(z,t)) = 2S_c(z) - \left(\sigma_c^{abs} v\right) n_c(z,t) \tag{67}$$

It is a linear reaction-convection-diffusion equation of parabolic type for unknown $n_c(z,t)$, expressed in conservation form. Two boundary conditions at interfaces $z = 0$ and $z = L$ must be given in order for the problem to be well-posed. Following the discussion presented in Section 3, we use the following conditions that depend on the charge c.

– at $z = 0$

$$\begin{aligned} -D_e \frac{\partial n_e(z,t)}{\partial z} - \mu_e n_e(z,t) E(z,t)\Big|_{z=0} &= -\frac{\kappa}{2-\kappa} v_e n_e(0,t), \\ -D_h \frac{\partial n_h(z,t)}{\partial z} + \mu_h n_h(z,t) E(z,t)\Big|_{z=0} &= 0. \end{aligned}$$

– at $z = L$

$$\begin{aligned} -D_e \frac{\partial n_e(z,t)}{\partial z} - \mu_e n_e(z,t) E(z,t)\Big|_{z=L} &= v_e n_e(L,t), \\ -D_h \frac{\partial n_h(z,t)}{\partial z} + \mu_h n_h(z,t) E(z,t)\Big|_{z=L} &= v_h n_h(L,t) \end{aligned}$$

5.2 Numerical scheme

We discuss an implicit finite-volume scheme, on a non uniform spatial grid and focus the analysis on the discrete maximum principle fulfilled by the numerical scheme for this linear reaction-convection-diffusion equation. The number of free charges $n_c(z,t)$ is positive, a discretization of the equation must give positive values. We know that for convection diffusion equation, upwind schemes induce artificial numerical diffusion that can be monitored thanks to the local Peclet number. Moreover a boundary layer characterizes the modelling. A discrete maximum principle must be verified for the discretization of

$$\nabla.(n_c(z,t)\mu_c s_c E(z,t)) \tag{68}$$

To this end, the finite volume discrete approximation is given by the following expression, where to simplify the notations, we have defined $v_c(z,t) = \mu_c s_c E(z,t)$,

$$\int_{\Omega_i} \nabla.(n_c(z,t)\, v_c(z,t)) = n_{i+\frac{1}{2}}^{k+1} v_{i+\frac{1}{2}}^{k+1} - n_{i-\frac{1}{2}}^{k+1} v_{i-\frac{1}{2}}^{k+1}. \tag{69}$$

Taking into account the signs of $v_{i+\frac{1}{2}}$ leads to

– if $v_{i+\frac{1}{2}}^{k+1} \geq 0$, then $n_{i+\frac{1}{2}}^{k+1} = n_i^{k+1}$, while if $v_{i+\frac{1}{2}}^{k+1} \leq 0$, then $n_{i+\frac{1}{2}}^{k+1} = n_{i+1}^{k+1}$.

A discrete maximum principle is then easily established following the method described in section 3. Work is in progress to analyse the numerical results.

6. Numerical software sirena

We describe briefly the architecture of our numerical software sirena. It is a toolbox for the numerical solution either by a two-fluxes method or by a reaction-diffusion method of the *see* yield. It is written in C language and consists of distinct modules that compute, the initial mesh either uniform or geometrically refined near $z = 0$, and solve the electric field equation, trapped charge density equation, etc . Several specialized data structures are used. The visualization is possible thanks to scripts written for gnuplot software but also in VTK format for the animated visualization of time-dependent quantities. It has been compiled under windows xp with a free C compiler, DEV-CPP while under linux with gnu gcc. A typical run with an adequate refined mesh requires less than a minute on a standard laptop.

7. Conclusions and perspectives

In this book chapter, we have presented a modelling for the computation of the initial and transient true see yield following a *traditional* two-fluxes approach. We have stressed the discrete maximum principle property of the conservative finite-volume numerical discretization presented in this chapter. A new asymptotic expression for the initial true see yield was presented and discussed.
A new approach, in this field, based on a reaction-diffusion modelling was presented for both initial and transient computation of true see yield. As in the two-fluxes approach, we have analyzed the discrete maximum principle properties of the finite-volume discretization scheme and provided some numerical simulations.
Finally, a numerical software **sirena** freely available upon request was presented.
In the future, We plan to extend this reaction-diffusion approach in two-spatial dimensions in order to perform numerical simulations of charge trapping inside the material for focalized electron beam, because in this case, lateral and longitudinal distributions of electrons/holes are important. This requires the knowledge for the creation of free electron/holes inside the sample. There appears to be no expression for such term in the litterature, which has a pear-like shape according to some monte carlo computations. We plan to use such computations and curve-fitting in order to obtain a law that will be plugged into the 2D version of **sirena**.

8. References

Aoufi, A. & Damamme, G. (2009). Numerical computation of secondary electron emission yield by a two-fluxes method, *Proceedings of ICNAAM 2009*, AIP, Rethymno - Crete.

Aoufi, A. & Damamme, G. (n.d.). 1d numerical simulation of charge trapping in an insulator submitted to an electron beam irradiation
part i: Computation of the initial secondary electron emission yield, *Applied Mathematical Modelling* . `www.elsevier.org`.

Fitting, H.-J. (1974). *Phys. Stat. Sol. A.* 26: 525–535.

G. Damamme, C. L. & Reggi, A. D. (1997). *IEEE Trans.Dielec.Elect.Insul* Vol. 5(No. 4): 558–584.

H.-J. Fitting, H. G. & Wild, W. (1977). *Physical Status Solid (a)* (43): 185–190.

I.A.Glavatskikh, V. S. K. & Fitting, H.-J. (2001). 'self-consistent electrical charging of insulating layers and metal-insulator-semiconductor structures', *J. Appl. Phys* (89): 440–448.

Levy, L. (2002). *Metal Charging*, Cepadues Edition.

5

Control of Photon Storage Time in Photon Echoes using a Deshelving Process

Byoung S. Ham
School of Electrical Engineering, Inha University
S. Korea

1. Introduction

Since the first protocol of quantum algorithm was put forth in 1994 (Shor, 1994), quantum information processing has been intensively studied (Nielsen & Chuang, 2000). The quantum approach has benefits over the classical optical information processing in areas such as prime number factorization (Shor, 1994), data searching (Grover, 1996), and high-resolution lithography (Boto et al., 2000; Yablonovitch, 1999). Compared to conventional cryptography based on public key cryptosystem (RSA cryptosystem) using conventional computers, prime number factorization using quantum computers has demonstrated a potential for a formidable attack on existing cryptographic systems. Like conventional memory which serves in the information processing unit, such as a processing unit together with logic gates, quantum memory is also essential to quantum information and communications networks. For quantum communications via a classical optical channel, the longest communication distance a quantum light can be transmitted is determined by the sensitivity of optical detectors and a lossy classical channel such as an optical fibre. Based on current technologies, the longest distance a single photon can propagate through an optical fibre is about 100 km (Zbinden et al., 1998). This distance should limit applications of quantum information especially for long-distance quantum communications. To solve the limited photon transmission, a quantum repeater has been introduced for virtually unlimited transmission distance (Duan et al., 2001; Jiang et al., 2007; Simon et al., 2007; Waks et al., 2002). Quantum memory is an essential element for the entangled photon swapping in the quantum repeaters. Because quantum repeaters swap entangled photons shared by neighboring remote quantum nodes in a quantum network, and the quantum information must be kept coherently through the quantum network, the minimum storage time of quantum memory is determined by the longest transmission distance of the lossy optical channel. For tranceatlantic quantum communications, roughly a one-second or more storage time is required. So far, such a long photon storage has not been demonstrated, where conventional quantum memory protocols limit the storage time to spin phase decay time at most ($\leq 10^{-3}$ second).

Unlike classical memories, quantum memory must satisfy a coherent process. Since the first observation of coherent retrieval of a stored optical pulse in a Bose Einstein condensate using slow light (Liu et al., 2001), interest in quantum memories has increased in the last decade (Alexander et al., 2006; Afzelius et al., 2010; Chaneliere et al., 2005. Choi et al., 2008;

Ham, 1998; Ham 2009a; Ham, 2010a; Hedges et al., 2010; Hetet et al., 2008; Hosseini et al., 2009; Julsgaard et al., 2004; Kocharovskaya et al., 2001; Kraus et al., 2006; Liu et al., 2001; Moiseev & Kroll, 2001; Moiseev et al., 2003; Neumann et al., 2009; Nilsson & Kroll, 2005; Sangouard et al., 2007; Turukhin et al., 2002; Van der Wal, et al., 2003). Because temporal multimode storage capability is required for the quantum repeaters, a photon echo-type protocol has emerged as a best candidate. Unlike a single atom-based quantum memory, echo-type quantum memory has the advantage of using an ensemble of atoms, where a quantum light is efficiently absorbed by many atoms. This ensemble system also provides near perfect storage capability as well as inherent temporal multimode capability. Following the first observation of echo-type optical memory in a spin system (Hahn, 1950), photon echoes were intensively studied in the 1980s and 1990s for spatiotemporal ultrahigh-speed all-optical information processing. Unlike all-optical memories, retrieval efficiency in quantum memories must satisfy at least a two thirds level of fidelity. In this chapter photon echo type quantum memory protocols are reviewed and compared. The chapter is composed of the following sections. In section 2, photon echoes are reviewed as a background of modified echo-type quantum memories. Section 3 presents the advantages and disadvantages of several modified photon echoes for quantum memory protocols. In Section 4, an optical locking technique is introduced for an ultralong photon storage method that can be applied to long-distance quantum communications. Section 5 discusses a phase matching condition for optical locking applied to different photon echo protocols, solving a main drawback in conventional photon echoes. Section 6 presents conclusions.

2. Review of photon echoes

Like spin echoes (Hahn, 1950), photon echoes (Kurnit, et al., 1964) use optical inhomogeneity of an atomic ensemble. Figure 1 shows numerical simulations of a two-pulse photon echo in a two-level atomic system. The first pulse D in Fig. 1(b) interacting with a two-level optical system excites atoms onto the excited state |2>. For a visualization purpose of maximum coherence, the first pulse D is set at a $\pi/2$ pulse area, where the pulse area Φ is defined by: $\Phi = \int \Omega dt$, and Ω is the Rabi frequency. By the interaction of the first pulse D, atomic coherence is created between states |1> and |3>. A phase relaxation-dependent decoherence is inevitable in any optical system. Because the atoms are inhomogeneously broadened, randomly detuned atoms from the absorption linecentre cause a fast dephasing of sum coherence for the atomic system. Later but before each individual atom diphases completely, the second pulse R, whose pulse area is π, interacts with all atoms whose sum coherence is washed out, and inverts the system to rephase. The rephasing by the second pulse R results in a time reversal process, where initial coherence should be retrieved after the same elapse as taken with R. Here, the photon echo as a coherent burst has nothing to do with a population transfer process but relates only to coherent phase retrieval of all individual atoms. The retrieval efficiency degrades as a function of time due to the optical phase decay process as well as to optical population decay of the excited atoms. In general, the optical phase decay time in rare-earth doped solids is ~0.1 ms, which is too short to quantum repeaters (Macfarlane & Shelby, 1987). Another problem of the two-pulse photon echoes is the echo reabsorption by the noninteracted (or nonabsorbed) atoms along the propagation direction, common in an optical medium governed by Beer's law, where the number of atoms excited by the light

pulse D is exponentially reduced as a function of propagation distance inside the medium (Sangouard, N. et al., (2007). Because the retrieval efficiency is defined by the ratio of emitted photon echo intensity to the data intensity, an optically dense medium is needed for near 100% data photon absorption. This optically thick medium is, however, disadvantageous to the echo generation due to echo reabsorption. As a result of reabsorption, the observed photon echo efficiency or retrieval efficiency in most rare-earth doped solids is less than 1 %. Hence the original photon echo protocols cannot be adapted for a quantum memory protocol unless the reabsorption problem is solved.

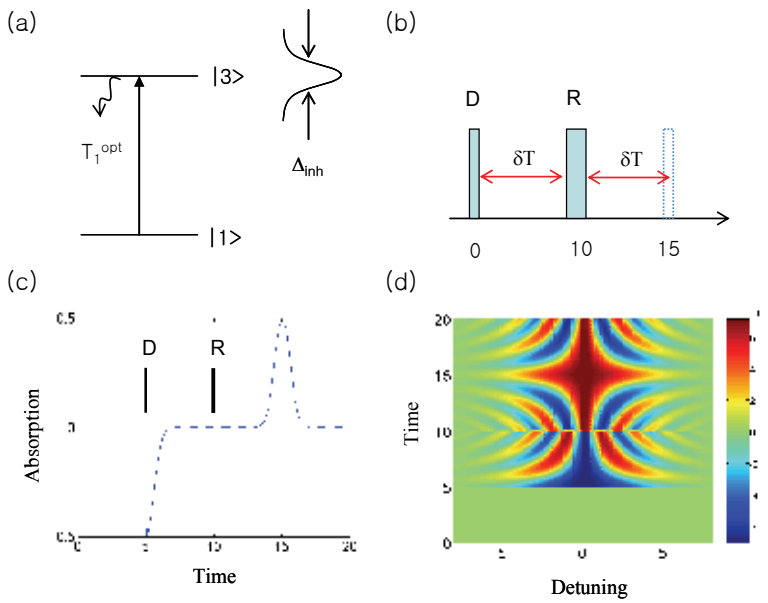


Fig. 1. Two-pulse photon echoes. (a) Energy level diagram interacting with light pulses, (b) pulse sequence for (a), (c) and (d) numerical simulations for (b). The pulse area of D and R is $\pi/2$ and π, respectively. All decay rates are assumed zero for visualization purposes. Optical inhomogeneous width Δ_{inh} is 680 kHz, where Rabi frequency of each pulse is 1 MHz

Compared with the two-pulse photon echoes (Kurnit, et al., 1964), a stimulated photon echo protocol was introduced to lengthen the storage time (Mossberg, 1982). In the stimulated photon echoes, the rephasing pulse R in the two-pulse photon scheme in Fig. 1(b) is divided into two $\pi/2$ pulses – that is W and R [see Fig. 2(a)]. By the first $\pi/2$ pulse, W, the atoms in both ground and excited states become spectrally modulated resulting in a spectral grating or frequency comb as shown in Fig. 2. Because the spectral modulation results from atom population modulation in the frequency domain caused by two consecutive optical pulses, D and W, the lifetime of the spectral grating is determined only by atom population decay time. Since the ground state population decay time is much longer than the optical counterpart, an optical deshelving technique to evacuate the excited atoms to a third state has been developed to increase the lifetime of the spectral grating (Mitsunaga & Uesugi, 1990). Thus, in the stimulated photon echoes, the storage mechanism is free from the optical phase decay process, which is the main storage mechanism to the two-pulse photon echoes. The third pulse R functions to rephase the coherence half-way stopped by W, resulting in a

stimulated echo (not shown). Here, the stimulated echo is a four-wave mixing process in the time domain, where R scatters off the spectral grating made by D and W, thus generating a time-delayed echo signal as shown in Fig. 3(a). The time delay of the echo from R is exactly the same as that between D and W due to the temporal four-wave mixing process. Thus, the storage time in the stimulated photon echoes can be eventually lengthened up to the spin population decay time, which is several orders of magnitude longer than the optical phase decay time (Macfarlane & Shelby, 1987). However, due to the excited state population loss during the storage process, the retrieval efficiency of the stimulated photon echoes must be less than 50%, which cannot satisfy the minimum fidelity of quantum memories.

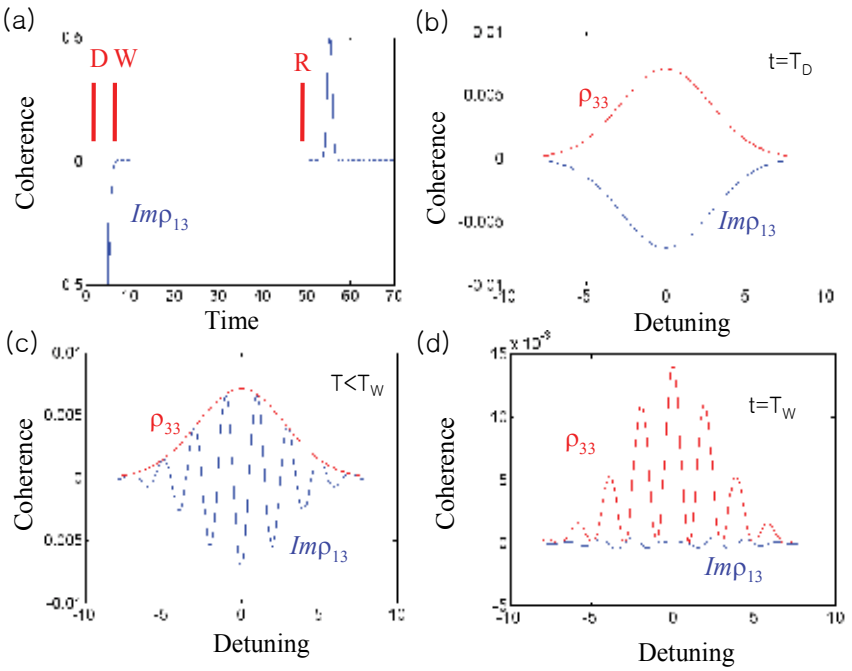


Fig. 2. Numerical simulations of stimulated photon echo. (a) ~ (d) sum of coherence $\mathrm{Im}\rho_{13}$ and excited state population ρ_{33}, where ρ_{ij} is a density matrix element defined by $\rho_{ij}=|i><j|$

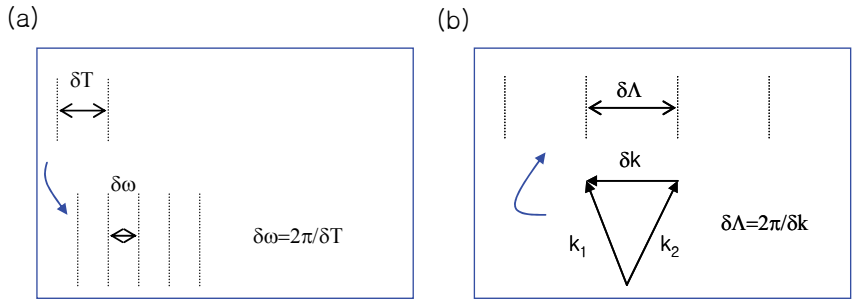


Fig. 3. Schematic diagram of (a) a spectral grating by D and W in Fig. 1(b) and (b) a spatial grating by two angled light beams k_1 and k_2

In both two-pulse and stimulated photon echoes, spontaneous emission noise due to population excitation should be a critical problem in quantum memory applications using single photons. The spontaneous emission noise problem, however, can be practically removed or alleviated if squeezed light or multiphoton entangled light (Marino et al., 2009) is used. Even in single photon-based quantum memory protocols, the spontaneous emission decay-caused quantum noise can be practically removed if an ultrashort pulse is used in a pencil-like geometry, where the pulse duration is still confined by optical inhomogenous width of the optical medium. Although Swiss and Calgary groups jointly criticised that photon echoes cannot be used for quantum memories due to the spontaneous emission noise, it fails with practical conditions in a rare-earth doped solids (Sangouard, N. et al., 2010).

In a rare-earth Pr^{3+} (0.05 at. %) doped Y_2SiO_5, which has been used for most modified photon echo based quantum memories (Afzelius et al., 2010, Ham, 2010d), total atom number per unit volume (cm^3) is $4.7x10^{18}$ (Maksimov et al., 1969). Either in the two-pulse photon echoes or in the stimulated photon echoes, at least one half the ground atoms are excited and spontaneously resulting in quantum noise. Thus, it seems obvious to say that even one out of 10^{18} atoms could affect the single photon-based echo signal to destroy the quantum fidelity. However, in a pencil-like propagation geometry, whose light cross section is 1 mm in diameter, the interaction volume decreases to 10^{-6} cm^3. For a 100 ps data pulse to cover a 4 GHz inhomogeneous width of the medium, the temporal ratio of the echo to the spontaneous decay time is 10^{-9}. Owing to the symmetry of echo to the data pulse in a virtual sphere made by a 10 cm focal length lens, the area ratio for the echo signal to the noise on the sphere is 10^{-5}. Thus, the effective number of spontaneously emitted photons affected to the echo signal is ~ 0.01. This number is nearly negligible to alter the photon echo fidelity.

3. Modified photon echoes for quantum memory applications

3.1 To solve the echo reabsorption problem in two-pulse photon echoes

Due to Beer's law, a trade-off exists between echo intensity and data absorption in an optically thick medium. If the echo propagation direction can be reversed to trace exactly along the data path, then no echo signals from the excited atoms interact with any nonexcited atoms due to the backward propagation scheme (Moiseev & Kroll, 2001). This idea has been experimentally demonstrated in 2009, where the echo enhancement factor even in an optically dilute medium is 15 times (Ham, 2009b). Another modified protocol to avoid echo reabsorption in the two-pulse photon echoes has been demonstrated by both a Lund group (Nilsson & Kroll, 2005) and Australian groups (Alexander et al., 2006; Hetet et al., 2008) using an electrical Stark effect. Instead of using π rephasing optical pulse, a pair of electrical stripe lines with opposing current flow spectrally controls the Stark effect, resulting in the same effect as the optical π rephsing pulse. Because the Rabi frequency of the electrical pulse is limited in most rare-earth doped solids, this electrical Stark method, however, limits the inhomogeneous width of atoms. Here, atom spectral width or inhomogeneous broadening determines the maximum amount of data, where the inverse of the spectral width determines the minimum pulse duration of the data D. Although echo efficiency can be maximized using this technique, the photon storage time is still limited by the optical phase decay time T_2^{opt} (in the order of 100 μs), which cannot satisfy the storage

time requirement for quantum repeaters (in the order of seconds) used for long-distance quantum communications.

3.2 To solve short storage time in two-pulse photon echoes

In a two-level system, the data pulse D excites optical coherence as mentioned in Fig. 1. Due to decoherence by optical phase decay time T_2^{opt}, however, individually excited coherence decreases as time elapses. Compared with optical coherence, spin coherence is much more robust, roughly ten times longer than the optical counterpart (Ham et al., 1997). Thus, if the optical coherence can be transferred into spin ensembles, longer storage time can be obtained (Moiseev & Kroll, 2001; Moiseev et al., 2003). In 1998, spin coherence excitation using temporally separated Raman optical pulses was investigated, where optical coherence between the optical pulses forming a Raman pulse plays a major role (Ham et al., 1998). The optical coherence in a time delayed Raman pulse is determined by inhomogeneous broadening of excited atoms. Contrary to general four-wave mixing processes, however, rephasing-based coherence transfer such as the stimulated photon echo is free from the optical coherence between control pulses. This will be discussed in more detail in Section 4.

3.3 To solve the spontaneous emission noise problem

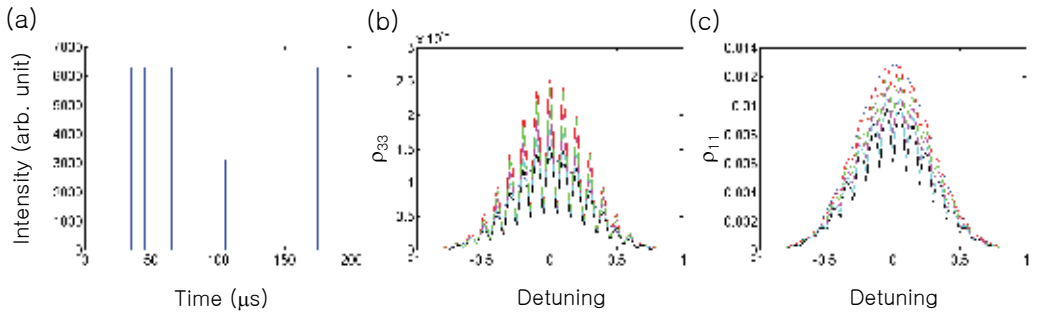


Fig. 4. Numerical simulations of AFC using five sets of two consecutive pulses. (a) pulse sequence, (b) excited state population, and (c) ground state population. Dotted: after the first pulse; Red: after the second pulse; Green: after the fourth pulse; Magenta: after the sixth pulse; Cyan: after the eight pulse; Black: after the tenth pulse in (a)

The spontaneous emission noise originates in the excited atoms due to optical population decay. Especially for quantum memories, the data pulse D must be weak, where only a small number of atoms are excited. By the rephasing pulse R, however, population inversion results in potential spontaneous emission noise. To solve this problem, an atomic frequency comb (AFC) method was introduced by a Swiss group (de Riedmatten et al., 2008). In AFC, the excited atoms are freely removed by a spontaneous emission decay process during atom preparation by a long optical train composed of two consecutive weak pulse pairs, as shown in Fig. 4(a). By the way, in the stimulated photon echoes, two consecutive optical pulses D and W in Fig. 2(a) create a spectral grating on both ground and excited states. If a $\pi/2$ optical pulse set is used, then ideally the spectral grating forms a 50% duty cycle with an equal distribution of atoms [see Fig. 2(d)]. In AFC, many weak-pulse sets accumulate to form one spectral grating on top of another to sharpen it, so that the increased finesse can be obtained as shown in Fig. 4. At the same time the excited state atoms freely decay down to a third

state. Eventually no excited state population remains. Thus, a spontaneous emission-free optical system can be achieved. Regarding the spectral grating, the physics of AFC for the retrieval process is exactly the same as for the stimulated photon echoes as discussed (Ham, 2010b). In AFC, however, a trade-off exists between high finesse and optical depth regarding enhanced retrieval efficiency. Even though the spectral grating can last up to spin population decay time, the storage time in AFC is determined by optical phase decay time T_2^{opt} (de Riedmatten et al., 2008), which is too short to quantum repeaters.

3.4 To solve the excited state population loss in stimulated photon echoes

This subsection somewhat overlaps with the modified two-pulse photon echoes in Section 3.2. In stimulated photon echoes, where longer storage time can be achieved, the excited atoms contain the same magnitude of coherence as the ground state. To avoid coherence loss due to optical population decay during the storage process, the excited atoms must be intentionally transferred into a third state for an on-demand halt, such as an auxiliary spin state. The coherence transfer technique suggested by the Lund group has been modified to transfer the spectrally modulated atoms from the excited state to the auxiliary spin state (Afzelius et al., 2010; Ham, 2009b). Because spectral grating is free from the optical phase decay process, storage time can be lengthened if population decay-caused coherence loss is halted. Unlike rephased atoms in the two-pulse photon echoes, the phase decay-dependent coherence loss can be frozen in the stimulated photon echoes using spectral grating, explaining why the coherence in AFC echoes (Afzelius et al., 2010) and phase locked echoes (Ham, 2009b) are degraded by spin dephasing. This will be discussed in more detail in Section 4.

4. Optical locking

As discussed in previous sections, a coherence transfer method is used to modify photon echoes to lengthen storage time. Here, an auxiliary deshelving pulse set (B1 and B2) is used to transfer atom population between the excited state (|3>) and an auxiliary spin state (|2>) as shown in Figs. 5(a) and (b). However, the atom population transfer between the optical and auxiliary spin states creates a π/2 phase gain, which applies evenly to all individual atoms assuming all light pulses are in phase. By a round-trip population transfer (by B1 and B2), the total phase gain accumulated becomes π. With the rephasing process serving to give a π phase shift to all individual atoms, this population transfer-based π phase gain completely washes out the rephasing performed by R leading to no echo generation. To avoid the odd phase gain obtained in the coherence transfer process, a phase recovery condition was investigated for an optical locking technique by an Inha group, S. Korea (Ham, 2009b). To create a multiple 2π phase shift during the coherence transfer process, the second deshelving pulse area must be 3π in order to make another round-trip population transfer for an additional π phase shift. As a general rule, the phase recovery condition of the optical deshelving pulses satisfies the followings (Ham, 2010a):

$$\Phi_{B1} = (4n-3)\pi, \tag{1}$$

$$\Phi_{B2} = (4n-1)\pi, \tag{2}$$

where n is an integer. Thus, the usage of an identical pulse set (Afzelius et al., 2010; Moiseev & Kroll, 2001) brings a contradiction, and violates the rephrasing process for echo

generation. The observation of delayed AFC echoes under this contradiction, however, can be explained as a result of coherence leakage due to imperfect population transfer in an optically dilute sample (Ham, 2010c; Ham, 2010d). The detected delayed echo signals in this case may be from the conventional photon echoes or at least mixed echoes. The remnant atom-generated echo signals from the excited state due to the imperfect population transfer cannot be separated from the delayed echoes (Ham, 2010c).

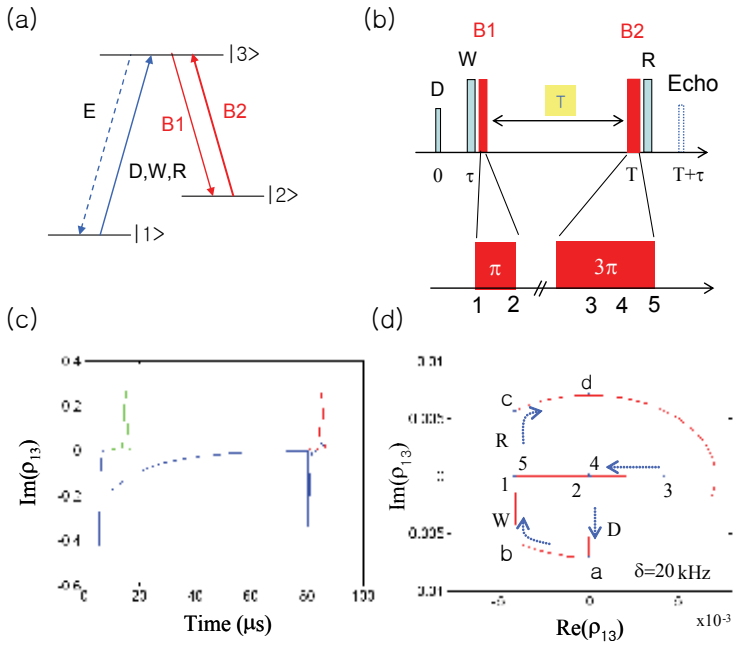


Fig. 5. Optical locking applied to stimulated photon echo. (a) An energy level diagram interacting with optical locking pulses, where D, W, and R represent DATA, WRITE, and READ pulses, respectively. (b) Optical pulse sequence for (a). (c) Numerical simulations. Red: for (b); Blue: without B1 and B2; Green: for two-pulse photon echo as a reference. (d) Bloch vector model without population decay loss. The numbers represent those in (b) indicating the phase recovery condition

In the experimental proofs of optical locking applied to both two-pulse photon echoes and stimulated photon echoes, the storage time extension of the photon echoes yields completely different results. First, in the two-pulse photon echoes, the storage time extension is governed by the overall spin dephasing rate, determined predominantly by spin inhomogeneous broadening, which is one tenth of the optical phase decay time (Afzelius et al., 2010; Ham, 2009b). As shown in the AFC and the phase locked echoes, this storage time extension is too short to solve the optical phase decay time constraint in conventional quantum memory protocols. With the stimulated photon echoes, however, the storage time extension is greatly increased due to the inherent property of optical phase locking resulting from the spectral grating based on population redistribution. The observed storage time extension, applying optical locking to the stimulated photon echoes in a rare-earth Pr^{3+} doped Y_2SiO_5 is five orders of magnitude longer than with the two-pulse photon echoes or AFC echoes (Ham, 2010d).

5. Phase matching in optical locking

As discussed in Section 3.1, a backward propagation technique has been introduced to solve the echo reabsorption problem. In two-pulse photon echoes using a rephasing halt, the phase matching condition for echo signal has nothing to do with the rephasing pulse vector, but relates with the data D and the optical locking pulses B1 and B2 (Ham, 2009b). Conversely, in the stimulated photon echoes in Section 3.4, the phase matching condition for echo signals includes the Data, Write, and Read pulses only, where optical locking pulses do not contribute at all to the phase matching condition (Ham, 2010d). From the results of these two cases, the important conclusion regarding the phase matching using optical locking is that the storage mechanism in each system is completely different. Thus, optical locking can result in an immense storage time extension in the stimulated photon echoes. For comparison, Fig. 6 represents a schematic of using optical locking pulses to these different cases. Only Fig. 6(b) can be applied to any meaningful storage time extension potential for quantum repeaters because of long photon storage time.

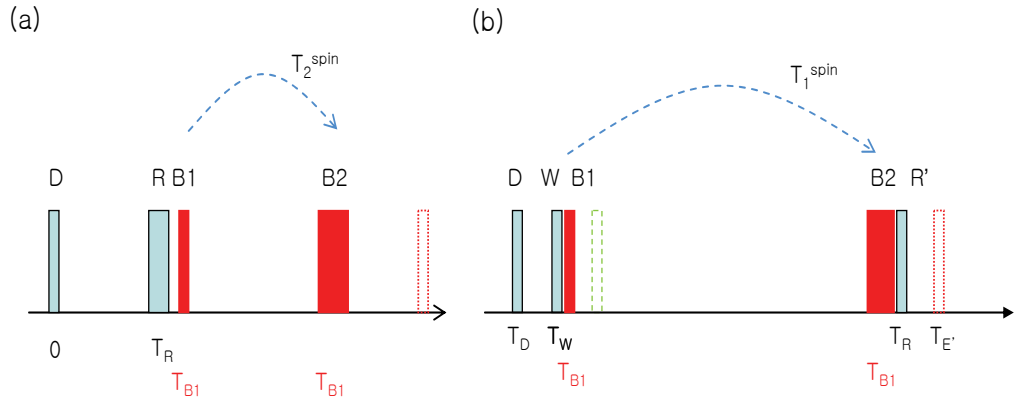


Fig. 6. Schematic diagram for (a) phase locked echo applied to rephased atoms in two-pulse photon echoes, and (b) optically locked echo applied to spectral grating in stimulated photon echoes. R and R′ represent for rephasing (2π) and READ (π) pulses, respectively

The physics of photon storage time extension in Fig. 6(b) is atom phase locking. This is accomplished by the spectral grating discussed in Fig. 2. This means that the phase grating excited by D is fully transferred into population grating by W, so that phase dependent decoherence is completely locked. This optical population information is coherently transferred into an auxiliary spin state |2> by B1 as optical-spin coherence conversion process in Fig. 5. In this stage, the spin dephasing becomes also an independent parameter to the coherence. Then, the last-long spin coherence is returned into state |3> by B2, and the optical population information is fully recovered into the optical phase grating by R′. In the experiment, a Korean group demonstrated one second storage time of photon echoes with 50% retrievcal efficiency (Ham, 2010d). Multi-photon entangled light or squeezed light could be the best candidate to this method. However, a single photon data pulse scheme can also be applied because of extremely low noise by the spontaneous emission decay process for a wide bandwidth, pencil-like propagation geometry as discussed above.

6. Conclusion

For potential applications of long-distance quantum communications using quantum memories, modified photon echo protocols have been reviewed. Although AFC echoes and gradient echoes have successfully solved the intrinsically low retrieval efficiency and spontaneous emission noise problems in the original photon echoes, the ultrashort photon storage time limits the usage to quantum memory applications. Instead, optical locking applied to the stimulated photon echoes has been demonstrated to prove ultralong photon storage time limited by spin population decay time, which is much longer than the minimum required storage time for quantum repeaters. Unlike critical objection by Swiss and Calgary group, the intrinsic atom population-caused spontaneous emission noise problem in the conventional photon echoes, however, can not be a serious problem due to low noise to the echo signal in a wide-bandwidth scheme. The key idea of storage time extension is locking phase decay process to the storage mechanism as well as optical-spin coherence transfer.

7. References

Afzelius, M. et al., (2010). Demonstration of atomic frequency comb memory for light with spin-wave storage. *Phys. Rev. Lett.* 104, 040503

Alexander, A. L. et al., (2006). Photon echoes produced by switching electric fields. *Phys. Rev. Lett.* 96, 043602

Boto et al., (2000). Quantum interferometric optical lithography: exploring entanglemnt to beat the diffraction limit. *Phys. Rev. Lett.* 85, 2733-2736

Choi, K. S. et al., (2008). Mapping photonic entanglement into and out of a quantum memory. *Nature* 452, 67-72

Chaneliere, T. A. et al., (2005). Storage and retrieval of single photons transmitted between remote quantum memories. *Nature* 438, 833-836

de Riedmatten, H. et al., (2008). Solid-state light-matter interface at the single-photon level. *Nature* 456, 773-777

Duan, L.-M. et al., (2001). Long-distance quantum communications with atomic ensembles and linear optics. *Nature* 414, 413-418

Grover, L. (1996). A fast quantum mechanical algorithm for database search," STOC '96, pp. 212-219.

Gurudev, M. V. et al. (2007).Quantum register based on individual electronic and nuclear spin qubits in diamond. *Science* 316, 1312-1316

Ham, B. S. et al., (1997). Frequency-selective time-domain optical data storage by electromagnetically induced transparency in a rare-earth doped solid. *Opt. Lett.* 22, 1849-1851

Ham, B. S. et al., (1998). Spin coherence excitation and rephrasing with optically shelved atoms. *Phys. Rev. B* 58, R11825-R11828

Ham, B. S. et al., (1999). Efficient phase conjugation via two-photon coherence in an optically dense crystal. *Phys. Rev. A* 59, R2583-R2586

Ham, B. S. (2009a). Ultralong quantum optical storage using reversible inhomogeneous spin ensembles. *Nature Photon.* 3, 518-522

Ham, B. S. (2009b). Phase locked photon echoes for near-perfect retrieval efficiency and extension storage time. arXiv:0911.3869

Ham, B. S. (2010a). Control of photon storage time using phase locking. *Opt. Exp*. 18, 1704-1713

Ham, B. S. (2010b). Analysis of controlled photon storage time using phase locking at atomic population transfer. arXiv:1004:0980

Ham, B. S. (2010c). A contradictory phenomenon of deshelving pulses in a dilute medium used for lengthened photon storage time. *Opt. Exp*. 18, 17749-17755

Ham, B. S. (2010d). On-demand control of photon echoes far exceeding the spin coherence constraint via coherence swapping between optical and spin transitions. arXiv:1010.4870.

Hahn, E. L. (1950). Spin echoes. *Phys. Rev*. 80, 580-594

Hedges, M. P. et al., (2010). Efficient quantum memory for light. *Nature* 465, 1052-1056

Hetet, G. et al., (2008). Elelctro-optic quantum memory for light using two-level atoms. *Phys. Rev. Lett.* 100, 023602

Hosseini, B. et al., (2009). Coherent optical pulse sequencer for quantum applications. *Nature* 461, 241-245

Jiang L. et al., (2007). Optical approach to quantum communications using dynamic programming. *PNAS* 104, 17291-17296.

Julsgaard, B. et al., (2004). Experimental demonstration of quantum memory for light. *Nature*, 432, 482-486

Kraus, B. et al., (2006). Quantum memory for nonstationary light fields based on controlled reversible inhomogeneous broadening. *Phys. Rev. A* 73, 020302

Kocharovskaya, O. et al., (2001). Stopping light via hot atoms. *Phys. Rev. Lett*. 86, 628-631

Kurnit, N. A. et al., (1964). Observation of a photon echo. *Phys. Rev. Lett.* 13, 567-568

Liu, C. et al., (2001). Observation of coherent optical information storage in an atomic medium using halted light pulses. *Nature* 409, 490-493

Macfarlane, R. M. & Shelby, R. M. (1987). *Coherent transient and holeburning spectroscopy of rare earth ions in solids.* Kaplyanskii, A. & Macfarlane, R. M., (Ed.) 30. Chap. 3. North-Holland

Marino, M. et al., (2009). Tunable delay of Einstein-Podolsky-Rosen entanglement. *Nature* 457, 859-862

Maksimov, B. A. et al., (1969). Crystal structure of Y-Oxysilicate Y2(SiO4)O. *Sov. Phys.-Doklady* 13, 1188-1190.

Mitsunaga, M. & Uesugi, N. (1990). 248-bit optical storage in Eu^{3+}:$YalO_3$ by accumulated photon echoes. *Opt. Lett.* 15, 195-197

Moiseev, S. A. & Kroll, S. (2001). Complete reconstruction of the quantum state of a single-photon wave packet absorbed by a Doppler-broadened transition. *Phys. Rev. Lett.* 87, 173601

Moiseev, S. A. et al., (2003). Quantum memory photon echo-like techniques in solids. *J. Opt. B: Quantum Semiclass. Opt.* 5, S497-S502

Mossberg, T. W. (1982). Time domain frequency-selective optical data storage. *Opt. Lett.* 7, 77-79

Neumann, P. et al., (2009). Ultralong spin coherence time in isotopically engineered diamond. *Nature Materials* 8, 383-387

Nielsen, M. A. & Chuang, I. L. (2000). *Quantum computation and quantum information*, Cambridge University Press, 0-521-63503-9, Cambridge, UK

Nilsson, M. & Kroll, S. (2005). Solid state quantum memory using complete absorption and re-emission of photons by tailored and externally controlled inhomogeneous absorption profiles. *Opt. Commun.* 247, 393-403

Sangouard, N. et al, (2010). Impossiblity of faithfully storing single phootns with the three-pulse photon echo. *Phys. Rev. A* 81, 062333.

Sangouard, N. et al., (2007). Analysis of a quantum memory for photon based on controlled reversible inhomogeneous broadening. *Phys. Rev. A* 75, 032327

Shor, P. (1994). Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer. *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, Santa Fe, NM, Nov. 20~22

Simon, C. et al., (2007). Quantum repeaters with photon pair sources and multimode memories. *Phys. Rev. Lett.* 98, 190503

Turukhin S. V., et al., (2002). Observation of ultraslow stored light pulses in a solid. *Phys. Rev. Lett.* 88, 023602

Van der Wal, C. H., et al., (2003). Atomic memory for correlated photon states. *Science* 301, 196-200

Waks, E. et al., (2002). Security of quantum key distribution with entangled photons against individual attacks. *Phys. Rev. A* 65, 052310

Yablonovitch, E. & Vrijen, R. (1999). Optical projedction lithography at half the Rayleigh resolution limit by two-photon exposure. *Opt. Eng.* 38, 334-338

Zbinden, H. et al., (1998). Quantum cryptography. *Appl. Phys. B* 67, 743-748

6

Waveguide Arrays for Optical Pulse-Shaping, Mode-Locking and Beam Combining

J. Nathan Kutz
Department of Applied Mathematics, University of Washington
USA

1. Introduction

Nonlinear mode-coupling (NLMC) is a well-established phenomenon which has been both experimentally verified (1; 2; 3; 4; 5) and theoretically characterized (6; 7; 8). NLMC has been an area of active research in all-optical switching and signal processing applications using wave-guide arrays (2; 3; 4; 5), dual-core fibers (1; 6; 7), and fiber arrays (9; 10). Recently, the temporal pulse shaping associated with NLMC has been theoretically proposed for the passive intensity-discrimination element in a mode-locked fiber laser (11; 12; 13; 14; 15; 16). The models derived to characterize the mode-locking consist of two governing equations: one for the fiber cavity and a second for the NLMC element (11; 12; 13; 16) (See Fig. 1). Although the two discrete components provide accurate physical models for the laser cavity, analytic methods for characterizing the underlying laser stability and dynamics is often rendered intractable. Thus, it is often helpful to construct an averaged approximation to the discrete components model in order to approximate and better understand the mode-locking behavior. Indeed, this is the essence of Haus' master mode-locking theory (17). Here, we develop an averaged approximation to the discrete laser cavity system based upon NLMC and characterize the resulting laser cavity dynamics. The resulting averaged equations are the equivalent of a master mode-locking theory for a laser cavity based upon nonlinear mode-coupling.

From an applications point of view, high-power pulsed lasers are an increasingly important technological innovation as their conjectured and envisioned applications have grown significantly over the past decade. Indeed, this promising photonic technology has a wide number of applications ranging from military devices and precision medical surgery to optical interconnection networks (17; 18; 19; 20). Such technologies have placed a premium on the engineering and optimization of mode-locked laser cavities that produce stable and robust high-power pulses. Thus the technological demand for novel techniques for producing and stabilizing high-power pulses has pushed mode-locked lasers to the forefront of commercially viable, nonlinear photonic devices. The performance of the waveguide array mode-locking model developed is optimized so as to produce high-power pulses in both the anomalous and normal dispersion regimes. The stability of the mode-locked solutions are completely characterized as a function of the cavity energy and the waveguide array parameters.

In principle, operation of a mode-locked laser (17; 18) is achieved using an *intensity discrimination* element in a laser cavity with bandwidth limited gain (17). The intensity

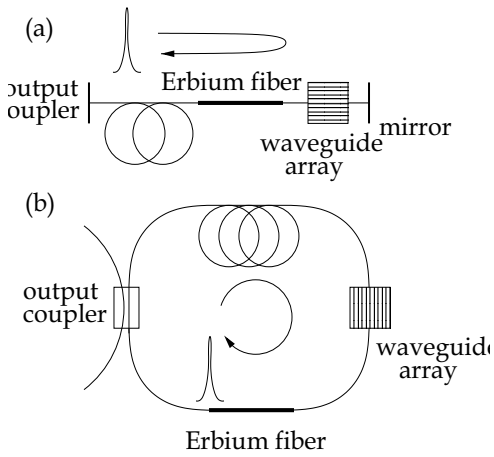


Fig. 1. Two possible laser cavity configurations which include nonlinear mode-coupling from the waveguide array as the mode-locking element. The fiber coupling in and out of the waveguide array occurs at the central waveguide as illustrated. Any electromagnetic field which is propagated into the neighboring waveguides is ejected (attenuated) from the laser cavity. In addition to the basic setup, polarization controllers, isolators, and other stabilization mechanisms may be useful or required for successful operation.

discrimination preferentially attenuates weaker intensity portions of individual pulses or electromagnetic energy. This attenuation is compensated by the saturable gain medium (e.g. Erbium-doped fiber). Pulse narrowing occurs since the peak of a pulse, for instance, experiences a higher net gain per round trip than its lower intensity wings. This pulse compression is limited by the bandwidth of the gain medium (typically ≈ 20 – 40 nm (17; 18)). It is well understood that some form of cavity saturable absorption or intensity discrimination is fundamental to producing stable mode-locked pulses in a passive laser cavity (17; 19; 23). Such intensity discrimination can be produced by a number of methods ranging from placing a linear polarizer in a fiber ring laser (24; 25; 26; 27), using a coupler in a figure-eight laser to produce nonlinear interferometry (28; 29; 30; 31), placing a semiconductor saturable absorber in a linearcavity configuration (32; 33; 34), or using a combination of spectral filtering with polarization filters in a dispersion controlled cavity (35; 36; 37; 38). Alternatively, active mode-locking can be used to produce mode-locked pulses by directly modulating the output electromagnetic field or using an acousto-optic modulator (40; 41). In all these cases, an effective intensity discrimination is generated to stabilize and control the mode-locked pulses. A relatively new method for generating intensity discrimination in a laser cavity is due to the nonlinear mode-coupling generated in a waveguide array (11; 12; 13; 21; 22). Although nonlinear mode-coupling has been proposed previously as a theoretical method for producing stable mode-locking (14; 16; 15), the waveguide array is the only nonlinear mode-coupling device that has been experimentally verified to produce the requisite pulse shaping required for mode-locking (42). This intensity discrimination, which is often only a small perturbation to the laser cavity dynamics, can be achieved with NLMC due to the well-known discrete self-focusing properties of the NLMC element. Indeed, the NLMC dynamics in wave-guide arrays is well-documented experimentally and provides the motivation for the current work. An overview of the techniques and methods which are capable of producing intensity discrimination and

mode-locking are reviewed in Refs. (17; 23). Although theoretical models have been developed towards understanding the mode-locking dynamics and stability of waveguide array based lasers (11; 12; 13; 21; 22), a characterization of its optimal performance and ability to generate high peak-power and high-energy pulses has not previously been performed.

Figure 1 illustrates two possible mode-locking configurations in which the waveguide array provides the critical effect of intensity discrimination (saturable absorption). In Fig. 1(a) a linear cavity configuration is considered whereas in Fig. 1(b) a ring cavity geometry is considered. In either case, the waveguide array provides an intensity dependent pulse shaping by coupling out low intensity wings to the neighboring waveguides. This low intensity field is then ejected from the laser cavity. In contrast, high intensity portions of the pulse are retained in the central waveguide due to self-focusing. Thus high intensities are only minimally attenuated. This intensity selection mechanism generates the necessary pulse shaping for producing stable mode-locked pulse trains.

2. Governing equations

In addition to the cavity (fiber) propagation equations, theoretical models are required to describe the NLMC element. Although nonlinear mode-coupling can be achieved in at least three ways (13) (wave-guide arrays, dual-core fibers, and fiber arrays), we will consider only wave-guide arrays since they illustrate all the basic properties of NLMC based mode-locking. The NLMC models are fundamentally the same, the only difference being in the number of modes coupled together. It should be noted that the NLMC theory presented here is an idealization of the dynamics of the full Maxwell's equations. For very short temporal pulses (i.e. tens of femtoseconds or less), modifications and corrections to the theory may be necessary.

2.1 Fiber propagation

The theoretical model for the dynamic evolution of electromagnetic energy in the laser cavity is composed of two components: the optical fiber and the NLMC element. The pulse propagation in a laser cavity is governed by the interaction of chromatic dispersion, self-phase modulation, linear attenuation, and bandwidth limited gain. The propagation is given by (17)

$$i\frac{\partial Q}{\partial Z}+\frac{1}{2}\frac{\partial^2 Q}{\partial T^2}+|Q|^2 Q+i\gamma Q-ig(Z)\left(1+\tau\frac{\partial^2}{\partial T^2}\right)Q=0, \quad (1)$$

where

$$g(Z)=\frac{2g_0}{1+\|Q\|^2/e_0}, \quad (2)$$

Q represents the electric field envelope normalized by the peak field power $|Q_0|^2$, and $\|Q\|^2=\int_{-\infty}^{\infty}|Q|^2\,dT$. Here the variable T represents the physical time in the rest frame of the pulse normalized by $T_0/1.76$ where T_0=200 fs is the typical full-width at half-maximum of the pulse. The variable Z is scaled on the dispersion length $Z_0=(2\pi c)/(\lambda_0^2\bar{D})(T_0/1.76)^2$ corresponding to an average cavity dispersion $\bar{D}\approx$12 ps / km-nm. This gives the one-soliton

peak field power $|Q_0|^2 = \lambda_0 A_{\text{eff}} / (4\pi n_2 Z_0)$. Further, n_2 = 2.6×10^{-16} cm^2/W is the nonlinear coefficient in the fiber, A_{eff} = 60 μm^2 is the effective cross-sectional area, λ_0 = 1.55 μm is the free-space wavelength, c is the speed of light, and $\gamma = \Gamma Z_0$ (Γ = 0.2 dB/km) is the fiber loss. The bandwidth limited gain in the fiber is incorporated through the dimensionless parameters g and $\tau = (1/\Omega^2)(1.76/T_0)^2$. For a gain bandwidth which can vary from $\Delta\lambda$=20–40 nm, $\Omega = (2\pi c / \lambda_0^2)\Delta\lambda$ so that $\tau \approx$ 0.08–0.32. The parameter τ controls the spectral gain bandwidth of the mode-locking process, limiting the pulse width.

It should be noted that a solid-state configuration can also be used to construct the laser cavity. As with optical fibers, the solid state components of the laser can be engineered to control the various physical effects associated with (1). Given the robustness of the mode-locking observed, the theoretical and computational predictions considered here are expected to hold for the solid-state setup. Indeed, the NLMC acts as an ideal saturable absorber and even large perturbations in the cavity parameters (e.g. dispersion-management, attenuation, polarization rotation, higher-order dispersion, etc.) do not destabilize the mode-locking.

2.2 Nonlinear mode-coupling equations

The leading-order equations governing the nearest-neighbor coupling of electromagnetic energy in the waveguide array is given by (2; 3; 4; 5; 8)

$$i\frac{dA_n}{d\xi} + C(A_{n-1} + A_{n+1}) + \beta |A_n|^2 A_n = 0, \tag{3}$$

where A_n represents the normalized amplitude in the n^{th} waveguide (n = –N, … ,–1, 0, 1, … , N and there are 2N + 1 waveguides). The peak field power is again normalized by $|Q_0|^2$ as in Eq. (1). Here, the variable ξ is scaled by the typical waveguide array length (4) of Z_0^*=6 mm. This gives $C = c Z_0^*$ and $\beta = (\gamma^* Z_0^* / \gamma Z_0)$. To make connection with a physically realizable waveguide array (5), we take the linear coupling coefficient to be c = 0.82 mm^{-1} and the nonlinear self-phase modulation parameter to be γ^* = 3.6 m^{-1}W^{-1}. Note that for the fiber parameters considered, the nonlinear fiber parameter is $\gamma = 2\pi n_2/(\lambda_0 A_{\text{eff}})$=0.0017 m^{-1}W^{-1}. These physical values give C = 4.92 and β = 15.1. The periodic waveguide spacing is fixed so that the nearest-neighbor linear coupling dominates the interaction between waveguides. Over the distances of propagation considered here (e.g. Z_0^* = 6 mm), dispersion and linear attenuation can be ignored in the wave-guide array.

The values of the linear and nonlinear coupling parameters are based upon recent experiment (4). For alternative NLMC devices such as dual-core fibers or fiber arrays, these parameters can be changed substantially. Further, in the dual-core fiber case, only two wave-guides are coupled together so that the n=0 and n=1 are the only two modes present in the dynamic interaction. For fiber arrays, the hexagonal structure of the wave-guides couples an individual wave-guide to six of its nearest neighbors. Regardless of these model modifications, the basic NLMC dynamics remains qualitatively the same.

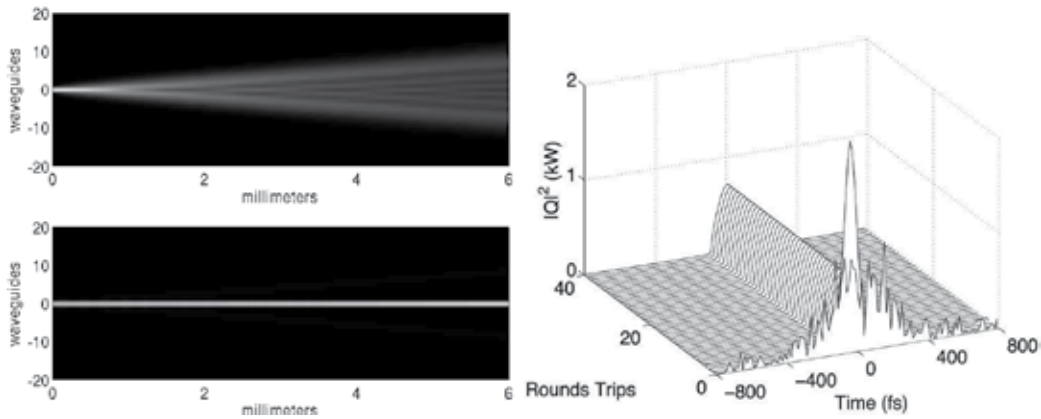
2.3 Mode-locking via NLMC

The self-focusing property of the wave-guide array is what allows the mode-locking to occur. The proto-typical example of the NLMC self-focusing as a function of input intensity is illustrated in Fig. 2a which is simulated with 41 (N = 20) waveguides (5) for two different

launch powers. For this simulation, light was launched in the center waveguide with initial amplitude $A_0(0) = 1$ (top) and $A_0(0) = 3$ (bottom). Lower intensities are clearly diffracted via nearest-neighbor coupling whereas the higher intensities remain spatially localized due to self-focusing. The spatial self-focusing can be intuitively understood as a consequence of (3) being a second-order accurate, finite-difference discretization of the focusing nonlinear Schrödinger equation (8). This fundamental behavior has been extensively verified experimentally (2; 3; 4; 5).

When placed within an optical fiber cavity, the pulse shaping associated with Fig. 2a leads to robust and stable mode-locking behavior (11; 12; 13). The computational model considered in this subsection evolves (1) while periodically applying (3) every round trip of the laser cavity (See Fig. 1). The simulations assume a cavity length of 5 m and a gain bandwidth of 25 nm ($\tau \approx 0.1$). The loss parameter is taken to be $\gamma = 0.1$ which accounts for losses due to the output coupler and fiber attenuation. To account for the significant butt-coupling losses between the waveguide array and the optical fiber, an additional loss is taken at the beginning and end of the waveguide array.

Figure 2b demonstrates the stable mode-locked pulse formation over 40 round trips of the laser cavity starting from noisy initial conditions with a coupling loss in and out of the waveguide array of 20% and with a constant gain $g_0 = 0.7$. Due to the excellent intensity discrimination properties of the waveguide array, the mode-locked laser converges extremely rapidly to the steady-state mode-locked solution. It is this generation of a stabilized mode-locked soliton pulse which the averaged model needs to reproduce. Note that the gain level g_0 has been chosen so that only a single pulse per round trip is supported. Further, in Fig. 2b the initial condition is chosen for convenience only.



(a) (b)

Fig. 2. (a) The classic representation of spatial diffraction and confinement of electromagnetic energy in a waveguide array considered by Peschel *et al.* (5). In the top figure, the intensity is not strong enough to produce self-focusing and confinement in the center waveguide, whereas the bottom figure shows the self-focusing due to the NLMC. Note that light was launched in the center waveguide with initial amplitude $a_0(0) = 1$ (top) and $a_0(0) = 3$ (bottom). (b) Stable mode-locking using a waveguide array with $g_0 = 0.7$. The mode-locking is robust to the specific gain level, cavity parameter changes, and cavity perturbations. Here is is assumed that a 20% coupling loss occurs at the input and output of the waveguide array due to butt-coupling (See Fig. 1b).

3. Pulse-shaping and X-waves in normal dispersion cavities

To illustrate the pulse shaping properties and the spontaneous formation of an X-wave structure in the normal GVD regime (43), we integrate numerically the proposed infinite-dimensional map by alternating Eqs. (1) and (2) for a length L_f and Eqs. (3) for a length L_a. Thus Q of Eq. (1) becomes A_0 in Eq. (3) when entering or leaving the waveguide array. Importantly, upon exiting the WGA, the system is strongly perturbed since the energy from all the neighboring channels (A_i where i = ±1, 2, 3, ...) are expelled from the laser cavity. Nevertheless, we observe the formation of a stable mode-locked pulse which shows the field A_0 at the output. The white-noise is quickly reshaped (over 10 round trips) into the mode-locking pulse of interest. Thus the mode-locking pulse acts as a *global attractor* to the laser cavity system. The simulation further implies that the mode-locking behavior is stable in the sense of Floquet (50) since it is a periodic solution in the cavity. The spectral shape clearly indicates that the mode-locking pulse is highly chirped, in analogy to what is found for 1D (no spatial dynamics) solutions of the master mode-locking equations in the normal GVD regime (17).

The overall electromagnetic field actually experiences a strong spatio-temporal reshaping per cavity round trip that involves stable coupling of a significant portion of the incoming WGA power to neighboring waveguides with nontrivial timing. The input and output time-domain intensities in all the waveguides, once nonlinear mode-locking has been achieved are displayed in Fig. 3. As shown, the interplay of accumulated GVD, discrete diffraction, and nonlinearity drives the field into a self-organized nonlinear X-waves, whose main signature is a central peak accompanied by pulse splitting occurring in the external channels. To show more clearly the X-shape of the mode-locking wave-packet generated at the ouput (B) of the waveguide array, Fig. 4 depicts a topographical plot of the time-domain (top) of all the waveguides. The distinctive X-wave structure is clearly evident. To lend further evidence to the existence of the X-wave structure, we plot the 2D Fourier transform of the time-domain. The right panel of Fig. 4 demonstrates that the spectrum is also X-shaped, as expected for X-waves (45; 46).

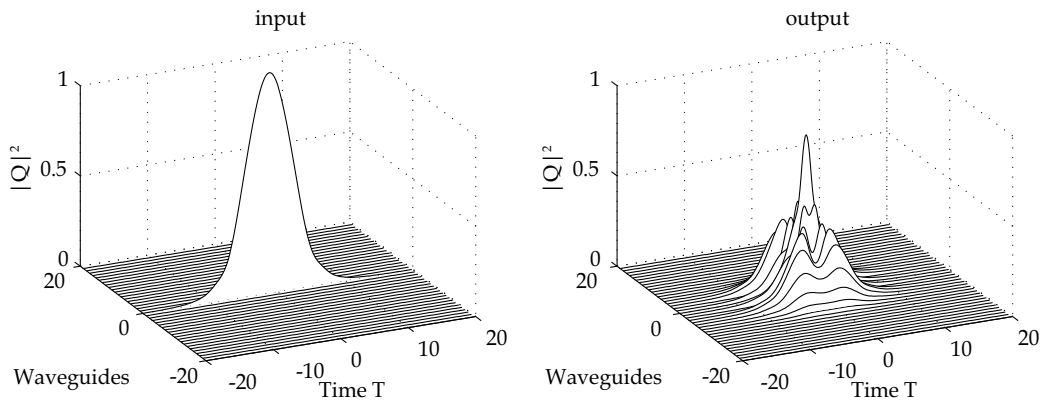


Fig. 3. Input (A) and output (B) temporal power distribution in the WGA. At the input, energy is only launched in the center waveguide (A_0), while at the output the energy has spontaneously formed into the X-wave configuration involving about eleven guides. Only energy in the A_0 mode is preserved upon re-entry into the fiber section of the cavity.

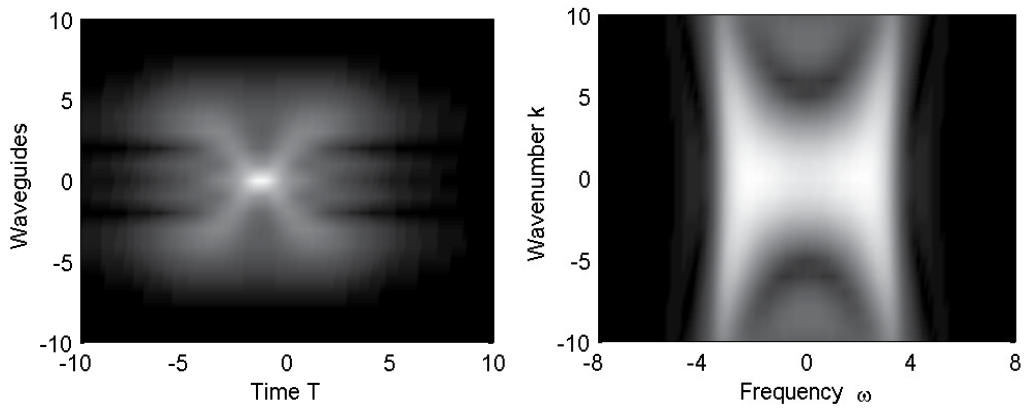


Fig. 4. Time-domain profiles and its two-dimensional Fourier transform at the output (B) in the WGA after steady-state mode-locking has been achieved. The X-wave structure is clearly seen in the topographical plot (top) of the output time-domain profiles of Fig. 3. Further, the expected wavenumber versus frequency dependence in the X-wave is shown in the Fourier domain (bottom).

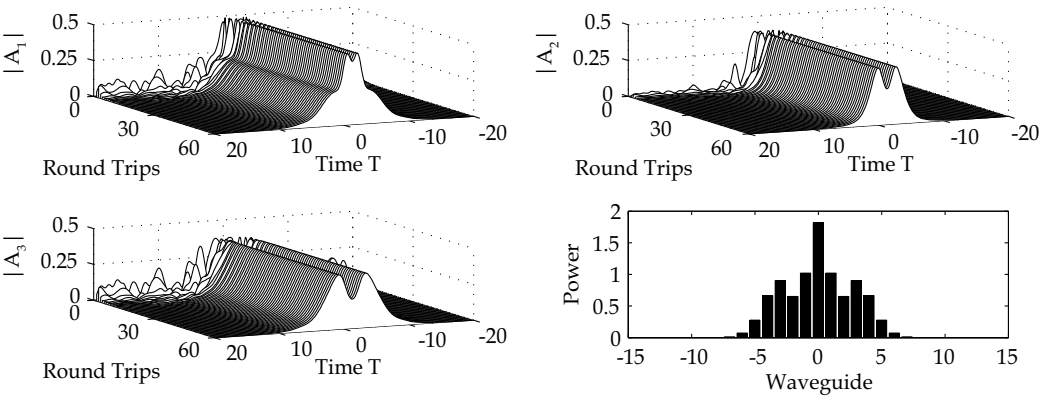


Fig. 5. Evolution to the steady-state output (B) in the neighboring waveguides A_1, A_2, and A_3. The bottom right graph is a bar graph of the steady-state distribution of energy $(\int_{-\infty}^{\infty} |A_j|^2 \, dT)$ in the waveguides. The symmetry about the center waveguide results from the initial condition being applied only in this waveguide. Note the significant re-distribution of energy in the waveguides.

To further characterize the mode-locking X-wave dynamics, Fig. 5 illustrates the mode-locking to the global attractor in the neighboring waveguides A_1, A_2 and A_3. Once again, generic white-noise initial data quickly self-organize into the steady-state mode-locking pattern. Note the characteristic pulse splitting (dip in the power) in the neighboring waveguides. This shows, in part, the generated X-wave structure. The final panel in Fig. 5 gives the energy $(\int_{-\infty}^{\infty} |A_j|^2 \, dT)$ in each of the waveguides and shows that a significant portion (more than 50%) of the electromagnetic energy has coupled to the neighboring waveguides. This is in sharp contrast to mode-locking with anomalous GVD for which less

than 6% is lost to the neighboring waveguides (13) and no stable X-waves are formed. The significant loss of energy in the cavity to the neighboring waveguides is compensated by the gain section and shows that the laser cavity is a strongly damped-driven system.

4. Averaged evolution models

The principle concept behind the averaging method presented here is to derive a single, self-consistent, and asymptotically correct representation of the dynamics in the laser cavity. In order to do so, we require an equation of evolution for each individual wave-guide which accounts for both the fiber propagation and wave-guide array coupling. Thus the term *averaged equations* refers to the governing set of equations which account for the average effect of dispersion, self-phase modulation, mode-coupling, attenuation, and bandwidth-limited gain in the wave-guide array based laser cavity configuration of Fig. 1. The following important guidelines must be met:

- Individual wave-guides are subject to chromatic dispersion and self-phase modulation.
- Coupling between neighboring wave-guides is a linear process with coupling coefficient C.
- The central wave-guide A_0 is subject to bandwidth-limited gain given in (1) and (2) since this wave-guide is coupled back into the fiber laser cavity. No other wave-guides experience gain due to amplification.
- The wave-guides neighboring the central wave-guide experience large attenuation due to the fact that they do not couple back into the laser cavity.

These simple guidelines, along with the governing equations (1), (2) and (3), allow for an asymptotically correct averaged description of the laser cavity dynamics.

Figure 6 shows a schematic of the averaging process which includes five wave-guides. Specifically, each wave-guide A_i is subject to two distinct physical propagation regions: the optical fiber region and the wave-guide array region. The period L of the laser cavity depicted theoretically in Fig. 6 is established with mirrors as demonstrated in Fig. 1. In the averaging process, only the center wave-guide A_0 experiences bandwidth-limited gain as given by (1) with (2) since this wave-guide contains the only optical fiber which has an Erbium doped section of fiber and physically butt-couples in and out of the wave-guide array (see Fig. 1). The optical fibers $A_{\pm}1$ and $A_{\pm}2$ representing the connection between wave-guide arrays are fictitious and only for averaging purposes. Indeed, as demonstrated in Fig. 1 the energy in the neighboring wave-guide arrays are allowed to escape the cavity into free-space. Put another way, one can think of the optical fiber propagation links in $A_{\pm}1$ and $A_{\pm}2$ in Fig. 6 as being governed by (1) with large attenuation but no gain, no dispersion, and no self-phase modulation. It should be noted that the attenuation in the neighboring wave-guides $A_{\pm}1$ may not be too large since the optical fiber radius is significantly larger than the wave-guide array diameters. Thus the butt-coupling process illustrated in Fig. 1b can transfer significant energy in $A_{\pm}1$ from one round-trip to the next.

The averaging is then accomplished by applying the principles of the split-step method, or Strang splitting, in reverse (44), i.e. we take the evolution for the two components of the laser cavity and fuse them into a single governing equation. In its simplest form, the split-step method decomposes a partial differential equation into two principle operators:

$$\frac{\partial A}{\partial Z} = N_1(A) + N_2(A) \tag{4}$$

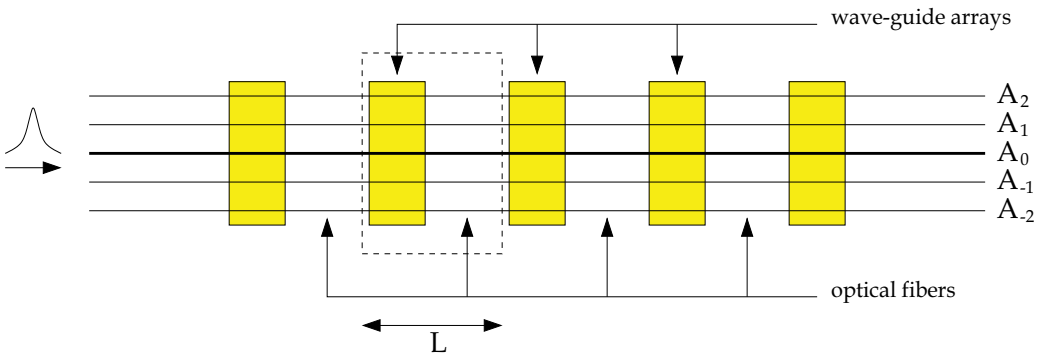


Fig. 6. Schematic of averaging process. Each wave-guide A_i is subject to two distinct physical propagation regions: the optical fiber region and the wave-guide array region. Here the period of the laser cavity L is determined by the mirror locations and fiber lengths in Fig. 1. The averaging procedure used is equivalent to the split-step method in reverse (44) which holds asymptotically for $L \ll 1$, i.e. a short cavity length.

where N_1 and N_2 are in general nonlinear operators which characterize two fundamentally different behaviors or phenomena (44). Here, N_1 and N_2 would represent the optical fiber propagation (1) and wave-guide array evolution (3) respectively. The split-step method then solves (4) numerically by decomposing it into two pieces over a single forward-step $\Delta Z \ll 1$:

$$\frac{\partial A}{\partial Z} = N_1(A) \tag{5a}$$

$$\frac{\partial A}{\partial Z} = N_2(A). \tag{5b}$$

Thus over each step ΔZ, the evolution is separated into two distinct evolution equations. Thus to advance the solution, (5a) would be solved for a ΔZ forward-step. The final solution of this step would be the initial data for (5b) which would also be advanced ΔZ. The two step process (5) is asymptotically equivalent to (4) provided the cavity period L, which is effectively ΔZ, is sufficiently small (44). The details of the split-step method and its asymptotic validity are outlined by Strang (44) and will not be considered here. In essence, the averaged equations account for the average dispersion, self-phase modulation, attenuation, gain and coupling which occurs over a single round trip of the laser cavity.
The only remaining modeling issue is the choice in the number of wave-guides ($n = 2N + 1$, see below (3)) to be considered in the averaged equations. From a practical viewpoint, each additional wave-guide considered implies the coupling of the system to another partial differential equation. Thus it is beneficial in the model to consider the minimal set of coupled equations which allow for the correct mode-locking dynamics. From a physical standpoint, the amount of energy in the wave-guides neighboring the central wave-guide is only a small fraction of the total cavity energy (13). This suggests that a small number of wave-guides can be considered.

4.1 Average cavity dynamics

When placed within an optical fiber cavity, the pulse shaping mechanism of the waveguide array leads to stable and robust mode-locking (11; 12; 13). In its most simple form, the

nonlinear mode-coupling is averaged into the laser cavity dynamics (21). Numerical simulations have shown that the fundamental behavior in the laser cavity does not change when considering more than five waveguides (21). It is interesting to note that if a three waveguide system is considered (one central and a neighboring waveguide on each side), mode-locking is not achieved. This can be explained due to the large attenuation required in the neighboring waveguides. This attenuation effectively reduces the coupling to the central waveguide which is critical for stable and robust mode-locking. Although it is not possible to consider the three waveguide model, further simplifications to the five waveguide model can be achieved by making use of the symmetric nature of the coupling and lower intensities in the neighboring waveguides (22). The resulting approximate evolution dynamics describing the waveguide array mode-locking is given by (22)

$$i\frac{\partial u}{\partial z}+\frac{D}{2}\frac{\partial^2 u}{\partial t^2}+\beta|u|^2 u+Cv+i\gamma_0 u-ig(z)\left(1+\tau\frac{\partial^2}{\partial t^2}\right)u=0 \tag{6a}$$

$$i\frac{\partial v}{\partial z}+C(w+u)+i\gamma_1 v=0 \tag{6b}$$

$$i\frac{\partial w}{\partial z}+Cv+i\gamma_2 w=0, \tag{6c}$$

where the $v(z, t)$ and $w(z, t)$ fields model the electromagnetic energy in the neighboring channels of the waveguide array. Note that the equations governing these neighboring fields are ordinary differential equations. All fiber propagation and gain effects occur in the central waveguide. It is this approximate system which will be the basis for our analytic findings. In fact, Eq. (6) provides a great deal of analytic insight due to its hyperbolic secant solutions

$$u(z,t)=\eta\ \mathrm{sech}\omega t^{1+iA}e^{i\theta z}, \tag{7}$$

where the solution amplitude η, width ω, chirp parameter A, and phase θ satisfy a set of nonlinear equations (22). Further, this solution forms from any arbitrary initial condition, thus acting as a global attractor to the system. This is in contrast to the master mode-locked equation (17) for which initial conditions must be carefully prepared to observe stable mode-locking.

In the anomalous dispersion regime ($D = 1 > 0$), solitonlike pulses can be formed as a result of the balance of anomalous dispersion and positive (i.e. self-focusing) nonlinearity. Typically mode-locked fiber lasers operating in the anomalous dispersion regime are limited in pulse energy by restrictions among the soliton parameters which is often referred to as the soliton area theorem (38). However, ultra-short, nearly transform-limited output pulses are desired for many applications. This encourages exploration of possible laser cavity configurations that could potentially maximize pulse energy in the anomalous dispersion regime. Figure 7 (left panel) shows the typical time- and spectral-domain mode-locking dynamics of the waveguide array model (6) in the anomalous dispersion regime. Here the equation parameters are $\beta = 8$, $C = 5$, $\gamma_0 = \gamma_1 = 0$, $\gamma_2 = 10$, $g_0 = 1.5$, and $e_0 = 1$. Stable and robust mode-locking is achieved from initial white-noise after $z \sim 100$ units. The steady state pulse solution has a short pulse duration and is nearly transform-limited, which is in agreement with experiments performed in the anomalous dispersion regime (17).

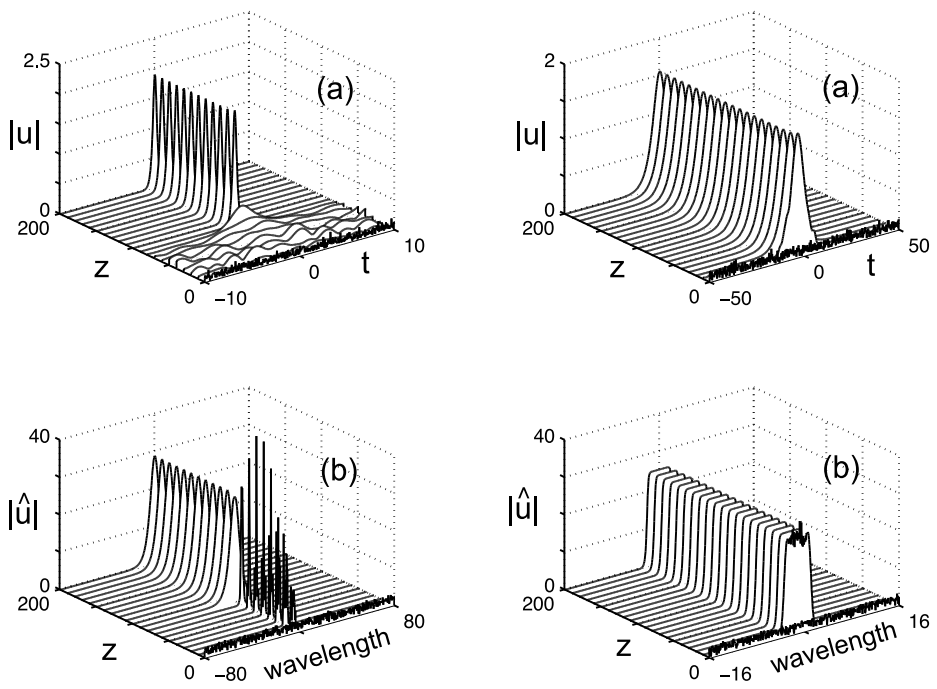


Fig. 7. Typical (a) time and (b) spectral mode-locking dynamics of the waveguide array mode-locking model Eq. (6) in the anomalous (left) and normal (right) dispersion regime from initial white-noise. For anomalous dispersion, the steady state solution is a short, nearly transform-limited pulse which acts as an attractor to the mode-locked system. For normal dispersion, the steady state solution is a broad, highly-chirped pulse which acts as an attractor to the mode-locked system.

Mode-locking in the normal dispersion regime ($D = -1 < 0$) relies on non-soliton processes and has been shown experimentally to have stable high-chirped, high-energy pulse solutions (35; 36). Figure 7 (right panel) shows the typical time and spectral mode-locking dynamics of the waveguide array model (6) in the normal dispersion regime. Here the equation parameters are $\beta = 1$, $C = 3$, $\gamma_0 = 0$, $\gamma_1 = 1$, $\gamma_2 = 10$, $g_0 = 10$, and $e_0 = 1$. In contrast to mode-locking in the anomalous dispersion regime, the mode-locked solution is quickly formed from initial white-noise after $z \sim 10$ units. The mode-locked pulse is broad in the time domain and has the squared-off spectral profile characteristic of a highly chirped pulse ($A \gg 1$). These characteristics are in agreement with observed experimental pulse solutions in the normal dispersion regime (33; 35; 36). Although these properties make the pulse solutions impractical for photonic applications, the potential for high-energy pulses from normal dispersion mode-locked lasers has generated a great deal of interest (38; 39; 47; 48).

5. Optimizing for high-power

As already demonstrated, the waveguide array provides an ideal intensity discrimination effect that generates stable and robust mode-locking in the anomalous and normal dispersion regimes. The aim here is to try to optimize (maximize) the energy and peak power output of the laser cavity. Intuitively, one can think of simply increasing the pump

energy supplied to the erbium amplifier in the laser cavity in order to increase the output peak power and energy. However, the mode-locked laser then simply undergoes a bifurcation to multi-pulse operation (22). Thus, for high-energy pulses, it becomes imperative to understand how to pump more energy into the cavity without inducing a multi-pulsing instability.

In what follows the stability of single pulse per round trip operation in the laser cavity is investigated as a function of the physically relevant control parameters. Two specific parameters that can be easily engineered are the coupling coefficient C and the loss parameter γ_1. Varying these two parameters demonstrates how the output peak power and energy can be greatly enhanced in both the normal and anomalous cavity dispersion regimes.

In order to assess the laser performance, the stability of the mode-locked solutions must be calculated. A standard way for determining stability is to calculate the spectrum of the linearization of the governing equations (6) about the exact mode-locked solution (7) (22; 49). The spectrum is composed of two components: the radiation modes and eigenvalues. The radiation modes are determined by the asymptotic background state where $(u,v,w) = (0, 0,0)$, whereas the eigenvalues are associated with the shape of the mode-locked solution (7). Details of the linear stability calculation and its associated spectrum are given in (22), while an explicit representation of the associated eigenvalue problem and its spectral content is given in (49). As in (22), a numerical continuation method is used here in conjunction with a spectral method for determining the spectrum of the linearized operator to produce both the solution curves and their associated stability. Our interest is in simultaneously finding stable solution curves and maximizing their associated output peak power and energy as a function of the parameters C and γ_1. In the normal and anomalous dispersion regimes, stable high peak power curves can be generated by increasing the input peak power via g_0. For the anomalous dispersion regime, while the peak power increase is a marginal ≈20%, the energy output can be doubled. For normal dispersion, the peak power increase is four fold with an order of magnitude increase in the output energy. These solutions then undergo a Hopf bifurcation before producing multi-pulse lasing (22).

Two types of instabilities are illustrated: (a) the instability of the bottom solution branch (dashed in Figs. 8(a)) when below the saddle node bifurcation point, and (b) the onset of the Hopf instability that leads to oscillatory, breathing solutions preceding the onset of the multi-pulsing instability. As is clearly demonstrated, the small-amplitude pulse below the saddle node bifurcation has one unstable eigenvalue whose eigenfunction is of approximately the form (7). This eigenfunction grows exponentially until the solution settles to the steady-state mode-locked solution. In contrast, the Hopf bifurcation generates two unstable modes at a prescribed frequency that leads to pulse oscillations (22).

5.1 Coupling coefficient *C*

To explore the laser cavity performance as a function of the coupling constant C, we consider the solution curves and their stability for a number of values of the coupling constant. Figure 8 shows the solution curves (η versus g_0) for both the anomalous and normal dispersion regimes as a function of the increasing coupling constant C. This figure demonstrates that an increased coupling constant allows for the possibility of increased peak power from the laser cavity. In the case of anomalous mode-locking, the peak power increase is only ≈15%, while for normal mode-locking the peak power is nearly doubled by

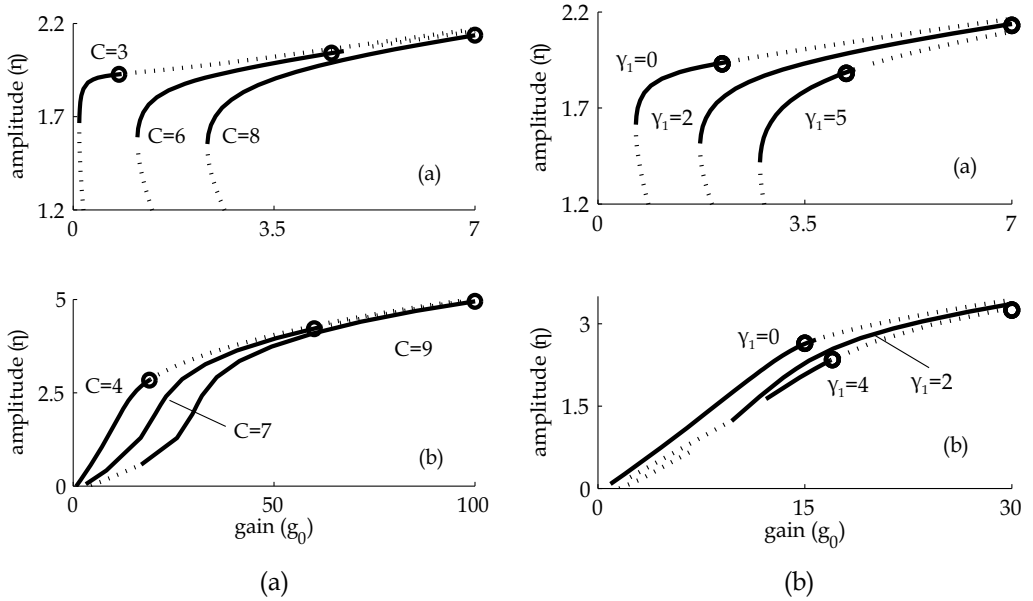


Fig. 8. Bifurcation structure of the mode-locked solution in the (a) anomalous and (b) normal dispersion regimes as a function of the coupling parameter C (left) and loss parameter γ_1 (right). The solid lines indicate stable solutions while the dotted lines represent the unstable solutions. For both anomalous and normal dispersion, an increase in the coupling constant leads to higher peak power pulses. The increase is ≈ 15% for anomalous dispersion and ≈ 100% for normal dispersion. There is also an optimal loss γ_1 for enhancing the output peak power by ≈ 25%. The circles approximately represent the highest peak power pulses possible for a given coupling or loss constant. The associated stable mode-locked pulse profile as a function of C is represented in Fig. 9 and the gain parameter is g_0 = 0.8, 4.5 and 7 for anomalous dispersion and g_0 = 19,60 and 100 for normal dispersion. The associated stable mode-locked pulse profile as a function of γ_1 is represented in Fig. 9 and the gain parameter is g_0 = 2.1,7 and 4.2 for anomalous dispersion and g_0 = 15,30 and 17 for normal dispersion.

increasing the linear coupling. The steady-state solution profiles are exhibited in Fig. 9 and verify the increased peak power associated with the increase in coupling constant C. Although the peak power is increased for the output pulse, it comes at the expense of requiring to pump the laser cavity with more gain. Although this makes intuitive sense, it should be recalled that the peak power and pulse energy levels are being increased without the transition to multi-pulse instabilities in the laser cavity.

5.2 Neighboring waveguide loss γ_1

To explore the laser cavity performance as a function of the loss constant γ_1, we consider the solution curves and their stability for a number of values of the loss constant. Figure 8 shows the solution curves (η versus g_0) for both the anomalous and normal dispersion regimes as a function of the increasing loss constant γ_1. This figure demonstrates that there is an optimal amount of loss in the neighboring waveguide that allows for the possibility of increased peak power from the laser cavity. In both the anomalous and normal cavities, the peak power increase is ≈ 25%. The steady-state solution profiles are exhibited in Fig. 9 and

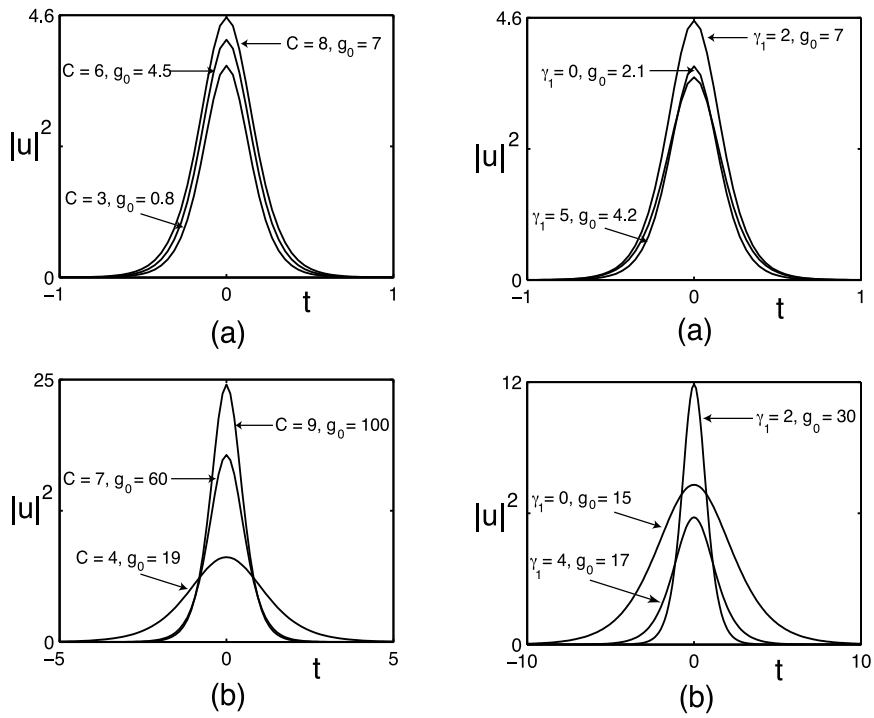


Fig. 9. Stable steady-state output pulse profiles for the (a) anomalous and (b) normal dispersion regimes corresponding to the circles in Fig. 8. As the coupling constant C increases (left figures), the peak power is increased ≈ 15% for anomalous dispersion and ≈ 100% for normal dispersion. The gain parameter is g_0 = 0.8, 4.5 and 7 for anomalous dispersion and g_0 = 19,60 and 100 for normal dispersion. As the loss constant γ_1 increases (right figures), the peak power is increased ≈ 25% for both anomalous and normal dispersion. The gain parameter is g_0 = 2.1,7 and 4.2 for anomalous dispersion and g_0 = 15,30 and 17 for normal dispersion.

verify the increased peak power associated with the increase in coupling constant γ_1. Although the peak power is increased for the output pulse, it comes at the expense of requiring to pump the laser cavity with more gain. Again recall that the peak power and energy levels are being increased without the transition to multi-pulse instabilities in the laser cavity.

5.3 Optimal design

Combining the above analysis of the mode-locking stability, we generate a three-dimensional surface representation of the stable mode-locking regimes. Figure 10 demonstrates the behavior of the stable solution curves as a function of g_0 (gain saturation parameter) versus C (coupling coefficient) versus $2\eta^2/\omega$ (the pulse energy). Both the anomalous and normal dispersion regimes are represented. Unlike Fig. 8, which represent the pulse intensities, here the pulse energy is represented and the pulse width parameter ω accounted for. Figure 10 (top) illustrates the stable solution curves for anomalous dispersion. It is clear that, as g_0 and C are increased, higher energy pulses can be achieved.

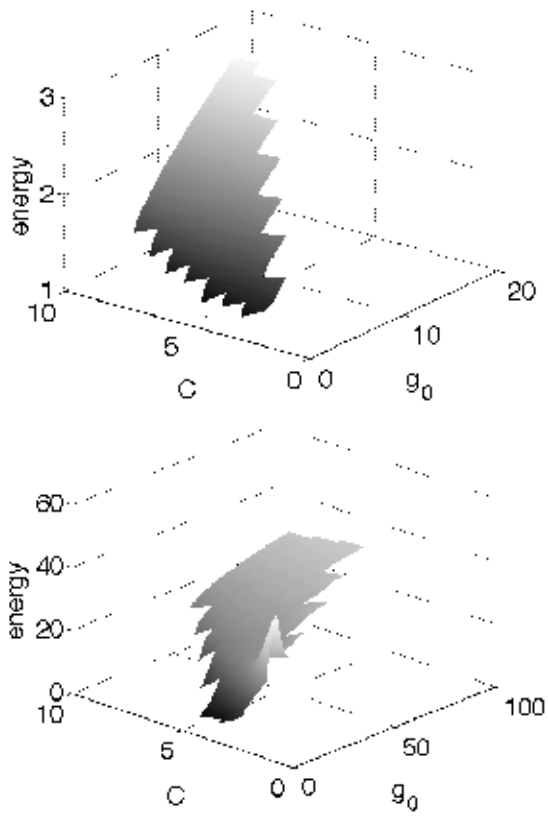


Fig. 10. The energy of stable mode-locked pulses is shown as a function of gain g_0 and coupling strength C for $\gamma_1 = 1.5$. Top is for anomalous and bottom is for normal dispersion. Note that by judiciously choosing the waveguide parameters, the energy output can be doubled in the anomalous regime and increased by an order of magnitude in the normal regime.

Indeed, the energy is nearly doubled for judicious choices of the parameters. Note that, although the energy is nearly doubled, the peak power only increases ≈20%. Likewise, Fig. 10 (bottom) illustrates the stable solution curves for normal dispersion. It is again clear that, as g_0 and C are increased, higher energy pulses can be achieved. In addition though, for low C values, there exists a small region of parameter space where high-energy pulses can be generated. However, the low C value, high-energy pulses are tremendously broad in the time domain and lose many of the technologically attractive and critical features of ultra-fast mode-locking.

6. Suppression of multi-pulsing for increased pulse energy

The onset of multi-pulsing as a function of increasing laser cavity energy is a well-known physical phenomenon (17; 23) that has been observed in a myriad of theoretical and experimental mode-locking studies in both passive and active laser cavities (51; 22; 52; 53; 54; 55; 56; 57; 58). One of the earliest theoretical descriptions of the multi-pulsing dynamics was by Namiki et al. (51) in which energy rate equations were derived for the averaged

cavity dynamics. More recently, a full stability analysis of the mode-locking solutions was performed showing that the transition dynamics between N and $N + 1$ pulses in the cavity exhibited a more complex and subtle behavior than previously suggested (22). Indeed, the theory predicted, and it has been confirmed experimentally since, that near the multi-pulsing transitions, both periodic and chaotic behavior could be observed as operating states of the laser cavity for a narrow range of parameter space (22; 52; 53). Here we generalize the energy rate equation approach to waveguide arrays (51) and develop an iterative technique that provides a simple geometrical description of the entire multi-pulsing transition behavior as a function of increasing cavity energy. The model captures all the key features observed in experiment, including the periodic and chaotic mode-locking regions (52), and it further provides valuable insight into laser cavity engineering for maximizing performance, i.e. enhancing the mode-locked pulse energy.

The multi-pulsing instability arises from the competition between the laser cavitie's bandwidth constraints and the energy quantization associated with the resulting mode-locked pulses, i.e. the so-called soliton area theorem (51). Specifically, as the cavity energy is increased, the resulting mode-locked pulse has an increasing peak power and spectral bandwidth. The increase in the mode-locked spectral bandwidth, however, reaches its limit once it is commensurate with the gain bandwidth of the cavity. Further increasing the cavity energy pushes the mode-locked pulse to an energetically unfavorable situation where the pulse spectrum exceeds the gain bandwidth, thereby incurring a spectral attenuation penalty. In contrast, by bifurcating to a two-pulse per round trip configuration, the pulse energy is then divided equally among two pulses whose spectral bandwidths are well contained within the gain bandwidth window.

6.1 Multi-pulsing transition

The basic mode-locking dynamics illustrated in Fig. 7 is altered once the gain parameter g_0 is increased. In particular, the analysis of the last section suggests that the steady-state pulse solution of Fig. 7 first undergoes a Hopf bifurcation before settling to a two pulse per round trip configuration. However, between the Hopf bifurcation and the stable two-pulse configuration there is a region of chaotic dynamics. Figure 11 shows a series of mode-locking behaviors which occur between the steady-state one pulse per round trip and the two pulses per round trip configurations. The gain values in this case are progressively increased from $g_0 = 2.3$ to $g_0 = 2.75$. As the dynamics change from one to two pulses per round trip steady-state, oscillatory and chaotic behaviors are observed. To characterize this behavior, we consider the gain dynamics $g(Z)$ of Eq. (2) in Fig. 12 which correspond to the evolution dynamics shown in Fig. 11. The gain dynamics provides a more easily quantifiable way of observing the transition phenomena.

At a gain value of $g_0 = 2.3$, the stable one-pulse configuration is observed in the top left panel of Fig. 11. The detailed evolution of this steady-state mode-locking process is shown in Fig. 7. The top right panel and middle left panel of Fig. 11 show the dynamics for gain values of $g_0 = 2.35$ and $g_0 = 2.5$ which are above the predicted threshold for a Hopf bifurcation. The resulting mode-locked pulse settles to a breather. Specifically, the amplitude and width oscillate in a periodic fashion. The oscillatory behavior is more precisely captured in Fig. 12 which clearly show the period and strength of oscillations generated in the gain $g(Z)$. Note that as the gain is increased further, the oscillations become stronger in amplitude and longer in period. To further demonstrate the behavior near the Hopf bifurcation, we

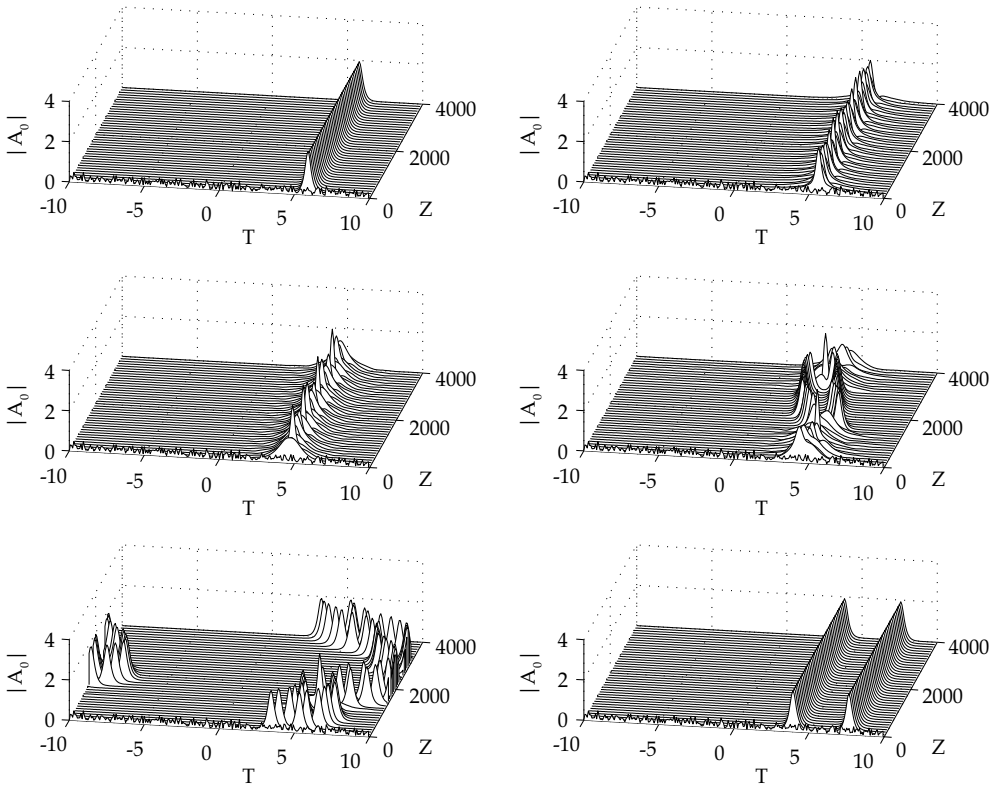


Fig. 11. Dynamic evolution and associated bifurcation structure of the transition from one pulse per round trip to two pulses per round trip. The corresponding values of gain are g_0 = 2.3, 2.35, 2.5, 2.55, 2.7, and 2.75. For the lowest gain value only a single pulse is present. The pulse then becomes a periodic breather before undergoing a "chaotic" transition between a breather and a two-pulse solution. Above a critical value ($g_0 \approx 2.75$), the two-pulse solution is stabilized. The corresponding gain dynamics is given in Fig. 12.

compute in Fig. 13 the Fourier spectrum of the oscillatory gain dynamics for g_0 = 2.35, 2.5 and 2.55. The dominant wavenumber of the Fourier modes for g_0 = 2.35 near onset is 10.07, which is in very good agreement with the theoretical prediction of 12.06 derived in Sec. 3.3 for the Hopf bifurcation. Increasing the gain further leads to an instability of the breather solution. The middle right and bottom left panels of Fig. 11, which have gain values of g_0 = 2.55 and g_0 = 2.7, illustrate the possible ensuing chaotic dynamics. Specifically, for a gain of g_0 = 2.55, the mode-locking behavior alternates between the breather and a two pulse per round trip state. The alternating between these two states occurs over thousands of units in Z. As the gain is further increased, the cavity is largely in the two-pulse per round trip operation with an occasional, and brief, switch back to a one-pulse per round trip configuration. Figure 12 illustrates the two chaotic behaviors in this case. Note the long periods of chaotic behavior for g_0 = 2.5 and the short bursts of chaotic behavior for g_0 = 2.7. Above g_0 = 2.75, the solution settles quickly to the two pulse per round trip configuration as shown in the bottom right panel of Fig. 11, which is therefore the new steady-state for the system. Thus the theoretical predictions of Sec. 3 capture the majority of the transition aside from the small window of parameter space for which the chaotic behavior is observed.

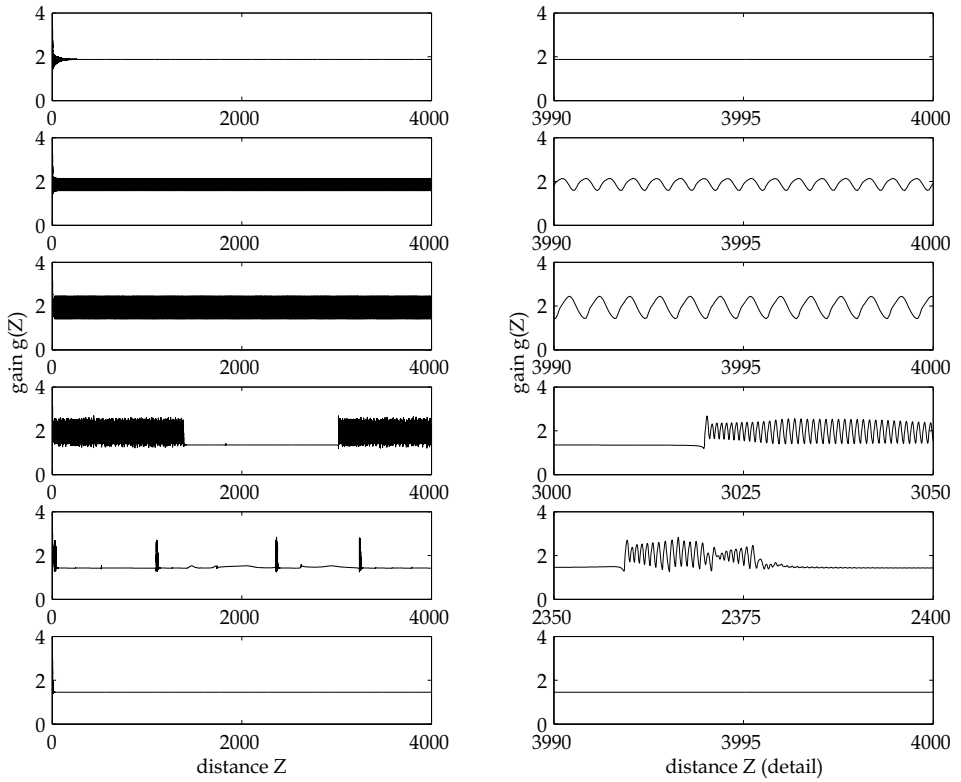


Fig. 12. Gain dynamics associated with the transition from one pulse per round trip to two pulses per round trip for the temporal dynamics given in Fig. 11. The left column is the full gain dynamics for $Z \in [0,4000]$, while the right column is a detail over $Z = 10$ or $Z = 50$ units, for values of gain equal to g_0 =2.3, 2.35, 2.5, 2.55, 2.7, and 2.75. Initially a single pulse is present (top panel), which becomes a periodic breather (following two panels) before undergoing a "chaotic" transition between a breather and a two-pulse solution (following two panels) until the two-pulse is stabilized (bottom panel) at $g_0 \approx 2.75$.

Bistability between the one- and two-pulse solutions in the laser cavity is easily demonstrated. The numerical simulations performed for this figure involve first increasing and then decreasing the bifurcation parameter g_0. Specifically, the initial value of g_0 =0.9 is chosen so that only the one-pulse solution exists and is stable. The value of g_0 is then increased to g_0 = 2.3 where the one-pulse solution is still stable. Increasing further to g_0 = 2.55 excites the Hopf bifurcation demonstrated in Figs. 11-13. Increasing to g_0 = 2.75 shows the two-pulse solution to be stable. The parameter g_0 is then systematically decreased to g_0 = 0.9, 2.3 and 2.55. Bistability is demonstrated by showing that at g_0 = 2.3 and g_0 = 2.55 both a one-pulse and two-pulse solution are stable. Dropping the gain back to g_0 = 0.9 reproduces the one-pulse solution shown in the top left panel. The top right panel shows the location on the solution curves (circles) where the one- and two-pulse solutions are both stable. It should be noted that the harmonic mode-locking is not just *bistable*. Rather, for a given value of the gain parameter g_0, it may be possible to have one-, two-, three-, four- or more pulse solutions all simultaneously stable. The most energetically favorable of these solution branches is the global-attractor of white-noise initial data.

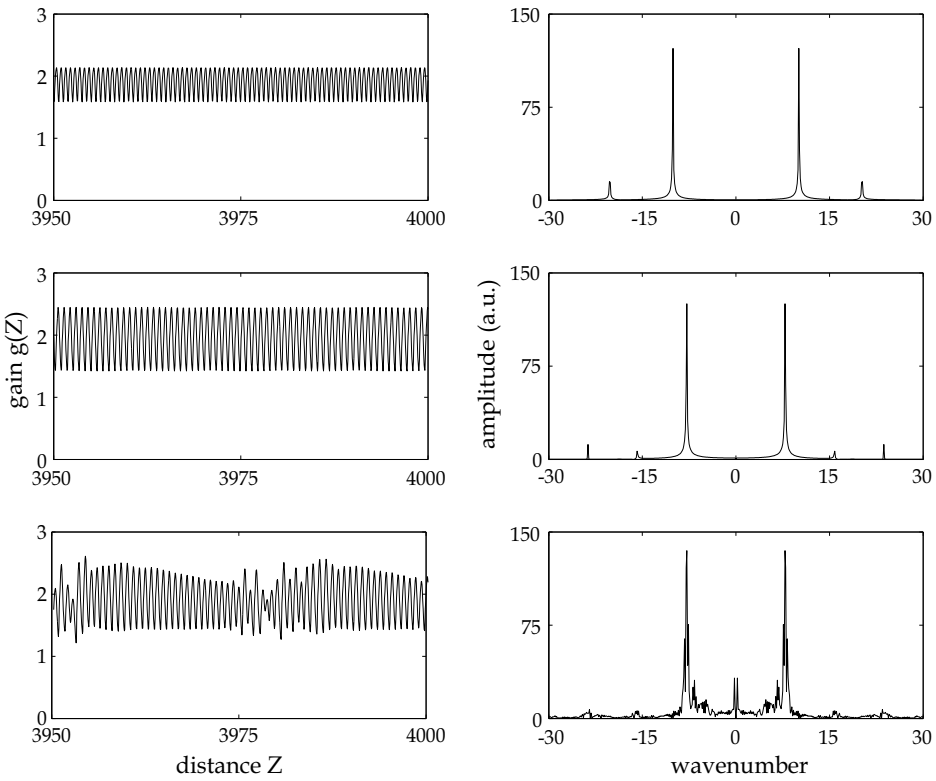


Fig. 13. Fourier spectrum of gain dynamics oscillations as a function of wavenumber. The last Z = 51.15 units (which gives 1024 data points) are selected to construct the time series of the gain dynamics and its Fourier transform minus its average. This is done for the data in Figs. 11 and 12 for g_0 = 2.35, 2.5 and 2.55.

6.2 Saturating gain dynamics

We will make the same assumptions as those laid out in Namiki et al. (51) and will simply consider a model for the saturating gain as well as the nonlinear cavity losses. The two primary components of loss and gain are included. The saturating gain dynamics results in the following differential equation for the gain (17; 23; 51):

$$\frac{dE_j}{dZ} = \frac{g_0}{1 + \sum_{j=1}^{N} E_j / E_{sat}} E_j \tag{8}$$

where E_j is the energy of the jth pulse (j = 1, 2, … N), g_0 measures the gain pumping strength, and E_{sat} is the saturation energy of the cavity. The total gain in the cavity can be controlled by adjusting the parameters g_0 or E_{sat}. In what follows here, the cavity energy will be increased by simply increasing the cavity saturation parameter E_{sat}. This increase in cavity gain can equivalently be controlled by adjusting g_0. These are generic physical parameters that are common to all laser cavities, but which can vary significantly from one cavity design to another. The parameter N is the number of potential pulses in the cavity (22). The mode-locked pulses are assumed to be identical as observed in both theory and experiment

(51; 22; 52; 53; 54; 55; 56; 57; 58). This parameter, which is critical in the following analysis, helps capture the saturation energy received by each individual pulse.

6.3 Nonlinear loss (saturable absorption)

The nonlinear loss in the cavity, i.e. the saturable absorption or saturation fluency curve, will be modeled by a simple transmission function:

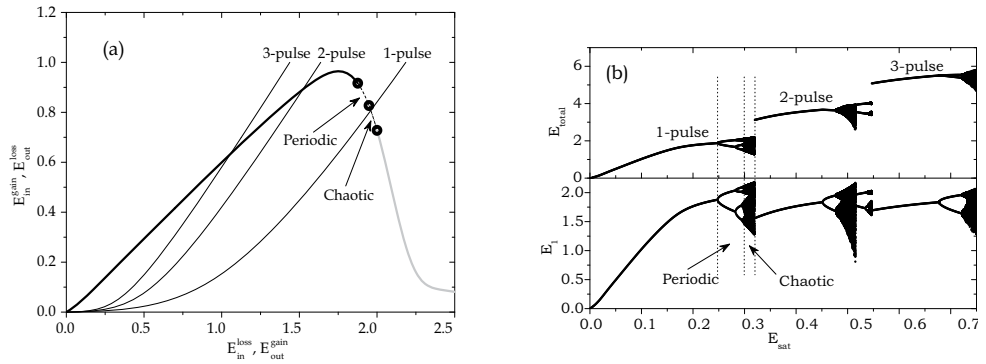
$$E_{out} = T(E_{in})E_{in}. \tag{9}$$

The actual form of the transmission function $T(E_{in})$ can vary significantly from experiment to experiment, especially for very high input energies. For instance, for mode-locking using nonlinear polarization rotation, the resulting transmission curve is known to generate a periodic structure at higher intensities. Alternatively, an idealized saturation fluency curve can be modified at high energies due to higher-order physical effects. As an example, in mode-locked cavities using wageguide arrays (22), the saturation fluency curve can turn over at high energies due to the effects of 3-photon absorption, for instance. Consider the rather generic saturation curve as displayed in Fig. 14. This shows the ratio of output to input energy as a function of the input energy. It is assumed, for illustrate purposes, that some higher-order nonlinear effects cause the saturation curve to turn over at high energies. This curve describes the nonlinear losses in the cavity as a function of increasing input energy for N mode-locked pulses. Also plotted in Fig. 14 is the analytically calculated terminus point which gives a threshold value for multi-pulsing operation. This line is calculated as follows: the amount of energy, E_{thresh}, needed to support an individual mode-locked pulse can be computed. Above a certain input energy, the excess amount of energy above that supporting the N pulses exceeds E_{thresh}. Thus any perturbation to the laser cavity can generate an addition pulse, giving a total of $N + 1$ pulses. This calculation, when going from $N = 0$ to $N = 1$, gives the self-starting threshold for mode-locking (51).

6.4 Iterative cavity dynamics

The generic loss curve along with the saturable gain as a function of the number of pulses Eq. (8) are the only two elements required to completely characterize the multi-pulsing transition dynamics and bifurcation. When considering the laser cavity, the alternating action of saturating gain and nonlinear loss produce an iteration map for which only pulses whose loss and gain balance are stabilized in the cavity. Specifically, the output of the gain is the input of the nonlinear loss and vice-versa. This is much like the logistic equation iterative mapping for which a rich set of dynamics can be observed with a simple nonlinearity (59; 60). Indeed, the behavior of the multi-pulsing system is qualitatively similar to the logistic map with steady-state, periodic and chaotic behavior all potentially observed in practice.

In addition to the connection with the logistic equation framework, two additional features are particular to our problem formulation. First, we have multiple branches of stable solutions, i.e. the 1-pulse, 2-pulse, 3-pulse, etc. Second the loss curve terminates due to the loss curve exceeding the threshold energy. Exhibited in this model are the input and output relationships for the gain and loss elements. Three gain curves are illustrated for Eq. (8) with $N = 1$, $N = 2$ and $N = 3$. These correspond to the 1-pulse, 2-pulse and 3-pulse per round trip solutions respectively. These curves intersect the loss curve that has been terminated at the threshold value. The intersection of the loss curve with a gain curve represents the mode-locked solutions. These two curves are the ones on which the iteration procedure occurs (59; 60).



(a) (b)

Fig. 14. (right) Nonlinear loss and saturating gain curves for a 1-pulse, 2-pulse and 3-pulse per round trip configuration. The intersection of the gain and loss curves represents the mode-locked solution states of interest. As the cavity energy is increased, the gain curves shift to the right. The 1-pulse solution first experiences periodic and chaotic behavior before ceasing to exist beyond the threshold point indicated by the right most bold circle. The solution then jumps to the next most energetically favorable configuration of 2-pulses per round trip. (Left) Iteration map dynamics for the nonlinear loss and saturating gain. Shown is the total cavity energy E_{out} (top panel) and the individual pulse energy E_1 (bottom panel) as a function of the cavity saturation energy E_{sat}. The transition dynamics between multi-pulse operation produces a discrete jump in the cavity energy. In this case, both periodic and chaotic dynamics are observed preceding the multi-pulsing transition. This is consistent with recent theoretical and experimental findings (22; 52).

Figure 14 (left) gives a quantitative description of the multi-pulsing phenomenon. Specifically, the nonlinear loss curve along with the gain curves of the 1-pulse, 2-pulse and 3-pulse mode-locked solutions are given along with the threshold point as before. As the cavity energy is increased through an increasing value of E_{sat}. The 1-pulse solution becomes unstable to the 2-pulse solution as expected. In this case, the computed threshold value does extend down the loss curve to where the periodic and chaotic branches of solutions occur, thus allowing for the observation of periodic and chaotic dynamics. The multi-pulsing bifurcation occurs as depicted in Fig. 14 (right). The total cavity energy along with the a single pulse's energy is depicted as a function of increasing gain. For this case, which is only a slight modification of the previous dynamics, the solution first undergoes a Hopf bifurcation to a periodic solution. Through a process of period doubling reminiscent of the logistic map (59; 60), the solution goes chaotic before eventually transitioning to the 2-pulse per solution branch. This process repeats itself with the transition from N to N + 1 pulses generating periodic and then chaotic behavior before the transition is complete. This curve is in complete agreement with recent experimental and theoretical findings (22; 52), thus validating the predicted dynamics.

The multi-pulsing instability ultimately is detrimental or undesirable for many applications where high-energy pulses are desired. Indeed, instead of achieving high-energy pulses as a consequence of increasing pump power, a multi-pulsing configuration is achieved with many pulses all of low energy. However, with the simple model presented here, it is easy to see that the laser cavity dynamics can be engineered simply by modifying the nonlinear loss

curve. Of course, modification of the loss curve is trivial do to in theory, but may be difficult to achieve in practice. Regardless, the potential for enhanced performance suggests that experimental modification of the nonlinear losses merit serious consideration and effort for the WGA. This essentially can circumvent the limitations on pulse energy imposed by the multi-pulsing instability. Thus the quantitative WGA cavity model can be used to pursue a more careful study of the nonlinear loss curves generated from physically realistic cavity parameters. Specific interest is in engineering the curve to increase performance before multi-pulsing occurs.

7. Beam combining with WGAs

Beam combining technologies are of increasing interest due to their ability to produce ultrahigh power and energy laser sources with standard, easy-to-implement fiber optic based laser cavities. The beam combining philosophy mitigates the nonlinear penalties (i.e. the multi-pulsing instability) that are incurred when attempting to achieve high power. By combining a large array of relatively low power cavities, each of which individually does not incur a nonlinear penalty, a high-power laser output can be produced. However, in order to be an effective technique, the laser beams need to be locked in phase. This has recently been accomplished in both active and passive continuous wave lasers (61; 62; 63; 64; 65). Indeed, a thorough review of the state of the field is in the 2009 special issue (61).

The waveguide array considered here is an ideal device for beam combining of pulsed, mode-locked lasers. Indeed, the nonlinear mode-coupling provided by the waveguide can be used to either combine mode-locked pulses directly in the waveguide, or combine laser cavities externally be running pulses through the WGA. Using the averaged cavity dynamics model (6), a generalized two-cavity model can be considered. Specifically, two linearly coupled cavities (6) are considered where the saturating gain in each cavity is independent of the other cavity. Interestingly enough, the two cavities shown in Fig. 15 can be engineered to be quite different by using different fiber segments and properties. Thus the dispersion, nonlinearity, gain, gain bandwidth and loss can lead to an unbalanced cavity. Alternatively, identical cavities can be constructed so that all the parameters (γ_1, β,D, e_0, g_{0j} ,C,τ) are approximately the same. Evidence for the ability of the cavity to beam combine pulses locked in time and phase is provided by Fig. 16. These are unpublished simulations which have only recently demonstrated the cavities ability to self-lock two cavities. The simulation parameters are D =1,C =1.23, and β = 7.3. Given how well the simulations have thus far proven to model the experimentally realizable cavity, it is expected that this highly promising result can be used as the basis for a pulsed, beam combining technology. Alternatively, the beam combining can be performed directly in the WGA itself in an appropriate parameter regime as demonstrated in Fig. 17. The parameters for the combining case are D = 0.5,C = 2.46 and β = 7.3, so that the coupling is relatively stronger then the time and phase locking considered in Fig. 16. In either configuration, the WGA can be used as an effective pulse combining technology. Engineering the cavity by changing key parameters such as the coupling allows us flexibility and full control of either post-cavity beam combining or intra-cavity beam combining. Such simulations, which represent recent state-of-the-art findings, are the first of their kind anywhere to demonstrate passive pulse combining. It clearly demonstrates the unique and promising role of WGA and supports the need for the current grant to explore such novel concepts and issues in mode-locking.

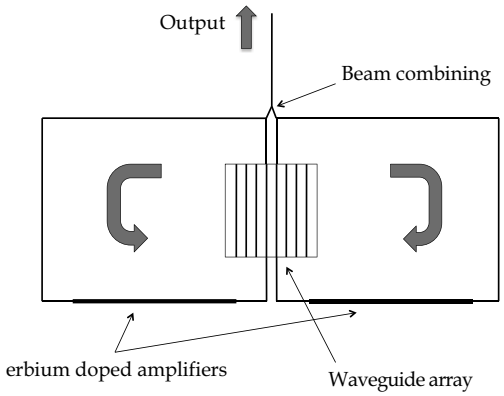


Fig. 15. Prototype for a mode-locked beam combining laser cavity. The WGA passively locks pulses in time and phase.

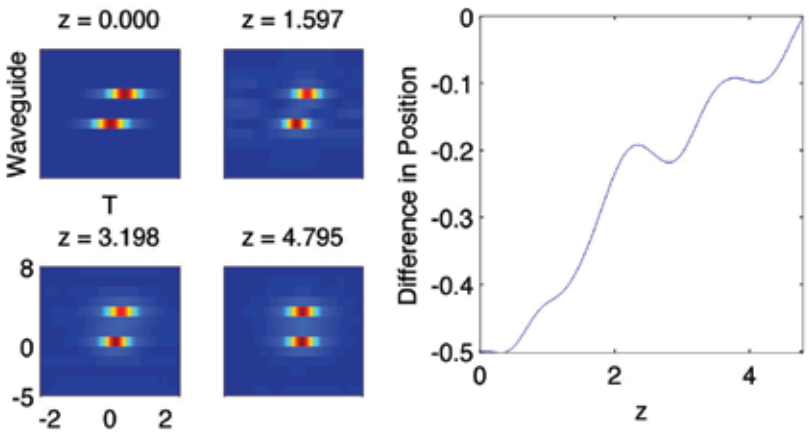


Fig. 16. Field intensities in the waveguide array for two identical initial pulses with a time-delay and a phase difference of $\pi/8$ (left), and the difference in pulse locations as a function of z (right). The pulses are attracted to each other as the system evolves.

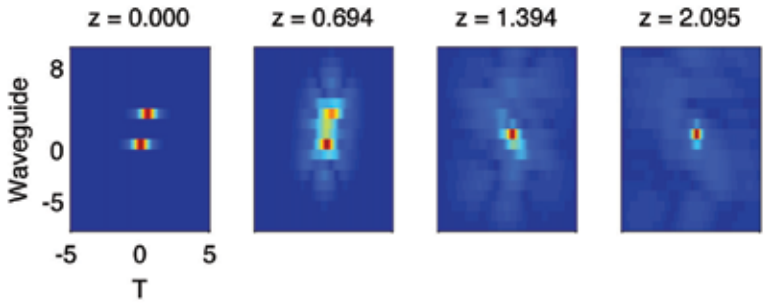


Fig. 17. Field intensitities in the waveguide for two identical pulses. Due to the relatively high coupling strength, the pulses are attracted towards each other across waveguides, eventually forming a single larger pulse. Note that by changing the initial separation of the pulses as well as the relative size of each pulse, the final location of the combined pulse can be controlled.

8. Conclusions

In conclusion, we have provided an extensive study of the robust and stable mode-locking that can be achieved by using the nonlinear mode-coupling in a waveguide array as the intensity discrimination (saturable absorption) element in a laser cavity. Indeed, the spatial self-focusing behavior which arises from the nonlinear mode-coupling of this mode-locking element gives the ideal intensity discrimination (or saturable absorption) required for temporal pulse shaping and mode-locking. Extensive numerical simulations of the laser cavity with a waveguide array show a remarkably robust mode-locking behavior. Specifically, the cavity parameters can be modified significantly, the coupling losses can be increased, and the gain model altered, and yet the mode-locking persists for a sufficiently high value of g_0. This demonstrates, in theory, the promising technological implementation of this device in an experiment. Here, the robust behavior as a function of the physical parameters is specifically investigated towards producing high peak-power and high-energy mode-locked pulses in both the normal and anomalous dispersion regimes.

In practice, the technology and components to construct a mode-locked laser based upon a waveguide array are available (5). An advantage of this technology is the short, nonlinear interaction region and robust intensity-discrimination (saturable absorption) provided by the waveguide array. The results here also suggest how the waveguide array spacing and waveguide array losses should be engineered so as to maximize the output intensity (peak power) and energy. Indeed, the theoretical framework established here provides solution curves and their stability as a function of the key parameters C and γ_1. These curves can be directly related to the design and optimization of laser cavities. A clear trend in anomalous mode-locking, normal mode-locking and spectrally filtered normal mode-locking is the increase in peak power and energy for the coupling coefficient C. For anomalous mode-locking this increase is only ≈25%, due to the soliton area theorem and the onset of multi-pulse instabilities. However, the pulse energy can be nearly doubled. For normal cavity fibers with or without filtering, the peak power increase can be four-fold with appropriate engineering, and the energy increase can be an order of magnitude. Thus the theory presented provides a critical design component for a physically realizable laser cavity based upon the waveguide array.

In the laser cavities proposed, index-matching materials, tapered couplers, polarization controllers and isolators may be useful and necessary to help further stabilize the theoretically idealized dynamics in the waveguide array model (6) presented here. Further, fiber tapering or free-space optics may be helpful to circumvent the losses incurred from coupling. Regardless, the theoretical results demonstrate that a mode-locked laser cavity operating by the nonlinear mode-coupling generated in a waveguide array is an excellent candidate for a compact, robust, cheap, and reliable high-energy pulse source based upon the union of the emerging technology of waveguide arrays with traditional fiber optical engineering.

9. Acknowledgements

J. N. Kutz is especially thankful to Brandon Bale, Matthew Williams, Edwin Ding, Colin McGrath, Frank Wise, Bjorn Sandstede, Steven Cundiff and Will Renninger for collaborative efforts on mode-locked laser theory in the past few years. J. N. Kutz is also acknowledges support from the National Science Foundation (NSF) (DMS-1007621) and the U.S. Air Force Ofce of Scientic Research (AFOSR) (FA9550-09-0174).

10. References

[1] Friberg, F.; Weiner, A.; Silberberg, Y.; Sfez, B.; & Smith, B. (1988). Femtosecond switching in a dualcore-fiber nonlinear coupler, *Opt. Lett.*, Vol. 13, 904-906.

[2] Eisenberg, H.; Silberberg, Y.; Morandotti, R.; Boyd, A. R.; & Aitchison, J. S. (1998). Discrete spatial optical solitons in waveguide arrays, *Phys. Rev. Lett.*, Vol. 81, 3383-3386.

[3] Aceves, A.; De Angelis, C.; Peschel, T.; Muschall, R.; Lederer, F.; Trillo, S.; & Wabnitz, S. (1996). Discrete self-trapping soliton interactions, and beam steering in nonlinear waveguide arrays, *Phys. Rev. E*, Vol. 53, 1172-1189.

[4] Eisenberg, H.; Morandotti, R.; Silberberg, Y.; Arnold, J.; Pennelli, G.; & Aitchison, J. (2002). Optical discrete solitons in waveguide arrays. 1. Soliton formation, *J. Opt. Soc. Am. B*, Vol. 19, 2938-1944.

[5] Peschel, U.; Morandotti, R.; Arnold, J.; Aitchison, J. S.; Eisenberg, H.; Silberberg, Y.; Pertsch, T.; & Lederer, F. (2002). Optical discrete solitons in waveguide arrays. 2. Dynamics properties, *J. Opt. Soc. Am. B*, Vol. 19, 2637-2644.

[6] Jensen, S. (1982). The nonlinear coherent coupler, *IEEE J. Quantum Electron.*, Vol. QE18, 1580-1583.

[7] Trillo, S. & Wabnitz, S. (1986). Nonlinear nonreciprocity in a coherent mismatched directional coupler, *App. Phys. Lett.*, Vol. 49, 752-754.

[8] Christodoulides, D. & Joseph, R. (1988). Discrete self-focusing in nonlinear arrays of coupled waveguides, *Opt. Lett.*, Vol. 13, 794-796.

[9] White, T.; McPhedran, R.; de Sterke, C. M.; Litchinitser, M.; & Eggleton, B. (2002). Resonance and scattering in microstructured optical fibers, *Opt. Lett.*, Vol. 27, 1977-1979.

[10] T. Pertsch, T.; Peschel, U.; Kobelke, J.; Schuster, K.; Bartelt, H.; Nolte, S.; Tünnnermann, A.; & Lederer, F. (2004). Nonlinearity and Disorder in Fiber Arrays, *Phys. Rev. Lett.*, Vol. 93, 053901.

[11] Kutz, J. N. (2005). Mode-Locking of Fiber Lasers via Nonlinear Mode-Coupling, In: *Dissipative Solitons*, N. N. Akhmediev and A. Ankiewicz, (Ed.) 241-265, Springer-Verlag, Berlin.

[12] Proctor, J. & Kutz, J. N. (2005). Theory and Simulation of Passive Mode-Locking with Waveguide Arrays, *Opt. Lett.*, Vol. 13, 2013-2015.

[13] Proctor, J. & Kutz, J. N. (2005). Nonlinear mode-coupling for passive mode-locking: application of wave-guide arrays, dual-core fibers, and/or fiber arrays, *Opt. Express*, Vol. 13, 8933-8950.

[14] Winful, H. & Walton, D. (1992). Passive mode locking through nonlinear coupling in a dual-core fiber laser, *Opt. Lett.*, Vol. 17, 1688-1690.

[15] Oh, Y.; Doty S.; Haus J.; & Fork, R. (1995). Robust operation of a dual-core fiber ring laser, *J. Opt. Soc. Amer. B*, Vol. 12, 2502-2507.

[16] K. Intrachat, K. & Kutz, J. N. (2003). Theory and simulation of passive mode-locking dynamics using a long period fiber grating, *IEEE J. Quant. Elec.* Vol. 39, 1572-1578.

[17] Haus, H. (2000). Mode-Locking of Lasers, *IEEE J. Sel. Top. Quant. Elec.* Vol. 6, 1173 1185.

[18] Duling, I. R. (1995). *Compact sources of ultrafast lasers*, Cambridge University Press, Cambridge, U. K.

[19] Keller, U. (2007). Ultrafast solid-state lasers, In: *Landolt Bernstein, LB VIII/1B*, Prof. Poprawe, (Ed.), Springer-Verlag, Berlin.

[20] Smith, K.; Davey, R.; Nelson, B. P.; & Greer, E. (1992). *Fiber and Solid-State Lasers*, Institution of Electrical Engineers, London.
[21] Proctor, J. & Kutz, J. N. (2007). Averaged models for passive mode-locking using nonlinear mode-coupling, *Math. Comp. in Sim.*, Vol. 74, 333-342.
[22] Kutz, J. N. & Sandstede, B. (2008) Theory of passive harmonic mode-locking using waveguide arrays, *Opt. Exp.*, Vol. 16, 636-650.
[23] Kutz, J. N. (2006) Mode-locked soliton lasers, *SIAM Rev.*, Vol. 48, 629-678.
[24] Tamura, K.; Haus, H.; & Ippen, E. (1992). Self-starting additive pulse mode-locked erbium fiber ring laser, *Elec. Lett.*, Vol. 28, 2226-2228.
[25] Haus, H.; Ippen, E.; & Tamura, K. (1994). Additive-pulse mode-locking in fiber lasers, *IEEE J. Quant. Elec.*, Vol. 30, 200-208.
[26] Fermann, M.; Andrejco, M.; Silverberg, Y.; & Stock, M. (1993). Passive mode-locking by using nonlinear polarization evolution in a polarizing-maintaining erbium-doped fiber laser, *Opt. Lett.*, Vol. 29, 447-449.
[27] Tang, D.; Man,W.; & Tam, H. Y. (1999). Stimulated soliton pulse formation and its mechanism in a passively mode-locked fibre soliton laser, *Opt. Comm.*, Vol. 165, 189-194.
[28] Duling, I. N. (1991) Subpicosecond all-fiber erbium laser, *Elec. Lett.*, Vol. 27, 544-545.
[29] Richardson, D. J.; Laming, R. I.; Payne, D. N.; Matsas, V. J.; & Phillips, M. W. (1991). Self-starting, passively mode-locked erbium fiber laser based on the amplifying Sagnac switch, *Elec. Lett.*, Vol. 27, 542-544.
[30] Dennis, M. L. & Duling, I. N. (1992). High repetition rate figure eight laser with extracavity feedback, *Elec. Lett.*, Vol. 28, 1894-1896.
[31] Ilday, F. Ö.; Wise, F. W.; & Sosnowski, T. (2002). High-energy femtosecond stretched-pulse fiber laser with a nonlinear optical loop mirror, *Opt. Lett.*, Vol. 27, 1531-1533.
[32] Kärtner, F. X. & Keller, U. (1995). Stabilization of solitonlike pulses with a slow saturable absorber, *Opt. Lett.*, Vol. 20, 16-18.
[33] Collings, B.; Tsuda, S.; Cundiff, S.; Kutz, J. N.; Koch, M.; Knox, W.; & Bergman, K. (1997). Short cavity Erbium/Ytterbium fiber lasers mode-locked with a saturable Bragg reflector, *IEEE J. Selec. Top. Quant. Elec.*, Vol. 3, 1065-1075.
[34] Tsuda, S.; Knox, W. H.; DeSouza, E. A.; Jan, W. J.; & Cunningham, J. E. (1995). Low-loss intracavity AlAs/AlGaAs saturable Bragg reflector for femtosecond mode-locking in solid-state lasers, *Opt. Lett.*, Vol. 20, 1406-1408.
[35] Proctor, B.; Westwig, E.; & Wise, F. W. (1993). Characterization of a Kerr-lens mode-locked Ti:sapphire laser with positive group-velocity dispersion, *Opt. Lett.*, Vol. 18, 1654- 1656.
[36] Haus, H. A.; Fujimoto, J. G.; & Ippen, E. P. (1991). Structures for additive pulse modelocking, *J. Opt. Soc. Am. B*, Vol. 8, 2068-2076.
[37] Ilday, F. O.; Buckley, J.; Wise, F. W.; & Clark, W. G. (2004). Self-similar evolution of parabolic pulses in a laser, *Phys. Rev. Lett.*, Vol. 92, 213902.
[38] Chong, A.; Buckley, J.; Renninger, W.; & Wise, F. (2006). All-normal-dispersion femtosecond fiber laser, *Opt. Express*, Vol. 14, 10095-10100.
[39] Chong, A.; Renninger, W. H.; & Wise, F. W. (2008). Properties of normal-dispersion femtosecond fiber lasers, *J. Opt. Soc. Am. B*, Vol. 25, 140-148.
[40] Kärtner, F. X.; Kopf, D.; & Keller, U. (1995). Solitary pulse stabilization and shortening in actively mode-locked lasers, *J. Opt. Soc. of Am. B*, Vol. 12, 486-496.

[41] Haus, H. A. (1975). A theory of forced mode locking, *IEEE J. Quant. Elec.*, Vol. 11, 323-330.

[42] Hudson, D.; Shish, K.; Schibli, T.; Kutz, J. N.; Christodoulides, D.; Morandotti, R.; & Cundiff, S. (2008). Nonlinear femtosecond pulse reshaping in waveguide arrays, *Opt. Lett.*, Vol. 33, 1440-1442.

[43] Kutz, J. N.; Conti, C.; & Trillo, S. (2007). Mode-locked X-wave lasers, *Optics Express*, Vol. 15, 16022-16028.

[44] Strang, G. (1968). On the construction and comparison of difference schemes, *SIAM J. Numer. Anal.*, Vol. 5, 506-517.

[45] Conti, C.;Trillo, S.; Di Trapani, P.; Valiulis, G.; Piskarskas, A.; Jedrkiewicz, O.; & Trull, J. (2003). "Nonlinear Electromagnetic X-waves", *Phys. Rev. Lett.*, Vol. 90, 170406.

[46] Di Trapani, P.; Valiulis, G.; Piskarskas, A.; Jedrkiewicz, O.; Trull, J.; Conti, C.; & Trillo, S. (2003). Spontaneously Generated X-shaped Light Bullets, *Phys. Rev. Lett.*, Vol. 91, 093904.

[47] Bale, B. G.; Kutz, J. N.; Chong, A.; Renninger, W. H.; & Wise, F. W. (2008). Spectral filtering for mode-locking in the normal dispersion regime, *Opt. Lett.*, Vol. 33, 931-934.

[48] Bale, B. G.; Kutz, J. N.; Chong, A.; Renninger, W. H.; & Wise, F. W. (2008). Spectral filtering for high-energy mode-locking in normal dispersion fiber lasers, J. Opt. Soc. Am. B, Vol. 25, 1763-1770.

[49] Bale, B. G.; Farnum, E.; & Kutz, J. N. (2009). Theory and Simulation of multi-frequency mode-locking using wave-guide arrays, *IEEE J. Quant. Elec.*, Vol. 44, 976-983.

[50] Guckenheimer, J. & Holmes, P. (1990). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields,* Springer-Verlag, Berlin.

[51] Namiki, S.; Ippen, E. P.; Haus, H. A.; & Yu, C. X. (1997). Energy rate equations for mode-locked lasers, Vol. 14, 2099-2111.

[52] Bale, B. G.; Kieu, K.; Kutz, J. N.; &Wise, F. (2009). Transition dynamics for multi-pulsing in mode-locked lasers, *Optics Express*, Vol. 17, 23137-23146.

[53] Xing, Q.; Chai, L.; Zhang, W.; & Wang, C. (1999). Regular, period-doubling, quasi-periodic, and chaotic behavior in a self-mode-locked Ti:sapphire laser, *Opt. Commun.*, Vol. 162, 71-74.

[54] Collings, B.; Berman, K.; & Knox,W. H. (1998). Stable multigigahertz pulse train formation in a short cavity passively harmonic modelocked Er/Yb ber laser, *Opt. Lett.*, Vol. 23, 123-125.

[55] Fermann, M. E. & Minelly, J. D. (1996). Cladding-pumped passive harmonically mode-locked ber laser, *Opt. Lett.*, Vol. 21, 970-972.

[56] Grudinin, A. B.; Richardson, D. J.; & Payne, D. N. (1992). Energy quantization in gure eight bre laser, *Electron. Lett.*, Vol. 28, 1391-1393.

[57] Guy, M. J.; Noske, P. U.; Boskovic, A.; & Taylor, J. R. (1994). Femtosecond soliton generation in a praseodymium uoride ber laser, *Opt. Lett.*, Vol. 19, 828-830.

[58] Davey, R. P.; Langford, N.; & Ferguson, A. I. (1991). Interacting solutions in erbium bre laser, *Electron. Lett.*, Vol. 27, 1257-1259.

[59] Devaney, R. (1989). *An Introduction to Chaotic Dynamical Systems*, 2nd ed., Addison-Wesley, Redwood City.

[60] Drazin, P. G. (1992). *Nonlinear systems*, Cambridge University Press, Cambridge, U. K.

[61] Leger, J.; Nilsson, J.; Huignard, J. P.; Napartovich, A.; & Shay, T. (2009). IEEE J. Sel. Top. Quantum Electron. special issue on *Laser Beam Combining and Fiber Laser Systems*, Vol. 15.

[62] Shay, T. (2006). Theory of electronically phased coherent beam combination without a reference beam, *Opt. Express*, Vol. 14, 12188-12195.

[63] Shay, T.; Benham, V.; Baker, J.; Sanchez, A.; Pilkington, D.; & Lu, C. (2007). Self-Synchronous and Self-Referenced Coherent Beam Combination for Large Optical Arrays, *IEEE J. Sel. Top. Quantum Electron.*, Vol. 13, 480-486.

[64] Wu, T.; Chang, W.; Galvanauskas, A.; & Winful, H. (2009). Model for passive coherent beam combining in fiber laser array, *Opt. Express*, Vol. 17, 19509-19518.

[65] Bochove, E. & Shakir, S. (2009). Analysis of a spatial-filtering passive fiber laser beam combining system, *IEEE J. Sel. Top. Quantum Electron.*, Vol. 15, 320-327.

7

Monte Carlo Methods to Numerically Simulate Signals Reflecting the Microvascular Perfusion

Figueiras Edite[1], Requicha Ferreira Luis F.[1],
De Mul Frits F.M.[2] and Humeau Anne[3]
[1]*Faculty of Sciences and Technology of Coimbra University, Physics Department, Instrumentation Center (GEI-CI), Coimbra,*
[2]*previously at University of Twente, Department of Applied Physics, Biomedical Optics Group, Enschede,*
[3]*Laboratoire d'Ingénierie des Systèmes Automatisés (LISA), University of Angers, Angers,*
[1]*Portugal*
[2]*The Netherlands*
[3]*France*

1. Introduction

Biomedical engineering can be defined as a part of the engineering domain that aims at better understanding biological systems and proposing new tools for diagnosis and therapeutic purposes. Among the activities of the biomedical field, one is dedicated to the analysis of the very small blood vessels (microcirculation). The study of the latter is important for diagnosis and follow-up of pathologies among which we can find diabetes. This chapter deals with Monte Carlo simulations applied to the microcirculation domain.
The microcirculation comprises the blood vessels of the most peripheral part of the vascular tree. Microcirculation includes capillaries, arterioles (small arteries), venules (small veins), and arteriovenous anastomosis (shunting vessels). In what follows, only skin microcirculation will be studied. Skin microcirculation is an important and complex system for thermoregulation, skin metabolism, and transcutaneous penetration. The monitoring of skin microcirculation can be useful to assess and to better understand skin physiology and diseases.
The skin microvascular network corresponds to different compartments. Thus, the epidermis is the top layer of the skin. Epidermis is avascular. Below the epidermis, the dermal papillae contains the capillaries. The latter are responsible for the exchange of oxygen and metabolites with the surrounding tissues. Therefore, the blood perfusion through capillaries corresponds to the nutritive blood flow. The deeper dermal structures contain the arterioles, venules, and shunting vessels. These vessels feed and drain the capillary network and aim at maintaining an adequate body temperature.
In order to monitor microvascular blood flow, the laser Doppler flowmetry (LDF) technique has been proposed in the 1970s (for a review see for example Öberg, 1990; Humeau et al., 2007; Rajan et al., 2009; Cal et al., 2010). LDF allows a non-invasive and real time monitoring

of the blood perfusion with a minimal influence in the parameters under study. It is applicable in experimental and in clinical settings.

In the LDF technique, a coherent light is brought to impinge on the tissues under study, generally through an optical fibre. The photons are scattered by the moving objects (mainly red blood cells) and by static structures. Scattering in moving objects modifies the direction and frequency of the photons according to the Doppler principle. Scattering in static structures only affects the direction of the photons. The remitted light is brought through another optical fibre to a photodetector where optical mixing of light shifted and unshifted in frequency gives rise to a stochastic photocurrent. The power spectral density of the latter depends on the number of red blood cells and their shape and velocity distribution within the scattering volume. The zero order moment scales with the concentration of moving red blood cells, provided the red blood cells concentration in tissue is low. The first moment of the photocurrent power spectrum scales with the product of the red blood cells concentration and average velocity (Bonner & Nossal, 1981).

In most LDF devices, the light source is a laser diode having a wavelength of 780 nm. However, other wavelengths can be used (450-800 nm). Within the visible range, the longer the wavelength, the deeper the transmission in tissue. In the range 600 to 1200 nm, the light penetrates deeply into tissue because of lower scattering and absorption coefficients. Very often the light is brought to and from the tissues under study by optical fibres. In the probe tip, the fibres are generally positioned with a core centre spacing of 250 to 500 μm. The photons migrate in the tissue in random pathways from the transmitting to the receiving fibres. When the distance between the fibre tips increases, the average path length of the detected photons increases, as well as the measurement depth (Jakobsson & Nilsson, 1993). Moreover, the average path length of the photons and the measurement depth depend on the optical properties of the tissue. Therefore, no comparison of perfusion values is possible between organs. That is why no absolute units are possible when using LDF.

The product of the average speed and concentration of moving blood cells in the scattering volume corresponds to the LDF signal and is generally referred as perfusion (see an example of LDF signal in Figure 1). Microvascular blood perfusion varies with time and from place to place (temporal and spatial variability). Therefore, and when using laser Doppler flowmeters based on optical fibres and thus for which the sampling volume is rather small, differences in perfusion readings appear when recording LDF signals from adjacent sites. This constitutes one of the main limitations of the LDF technique. Laser Doppler imagers have emerged providing a monitoring of the perfusion in two dimensions.

LDF can be used in experimental investigations and in clinical trials. Many organs can be investigated with the LDF technique: kidney, liver, intestines, brain, skin. Moreover, the clinical applications are numerous: diabetes microangiopathy, flap monitoring, peripheral vascular disease, plastic surgery, Raynaud's phenomenon, thermal injury (see for example Ray et al., 1999; Humeau et al., 2004; Ziegler et al., 2004; Yamamoto-Suganuma & Aso, 2009; Merz et al., 2010; Smit et al., 2010).

Due to the poor predictability of tissue perfusion for given location and time, LDF signals are rarely recorded in basal conditions. Instead, stimuli are often provoked and responses are studied to reveal possible malfunctioning of the microcirculation. All in all, the main advantages of the LDF technique are the following: non-invasiveness, continuous recordings, easy to use, strong theoretical basis. By contrast, the main drawbacks of the technique are: no absolute calibration, no possibility to distinguish between nutritive (capillary) perfusion and global tissue perfusion, no comparison possible between

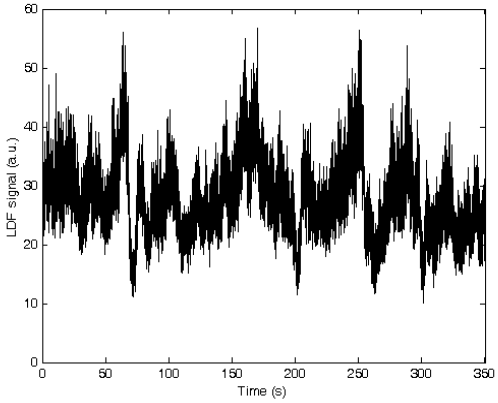


Fig. 1. Skin LDF signal recorded on the finger of a healthy subject at rest.

organs due to the variations of the photon path lengths because of the different optical properties of the tissues, variation between different individuals and in the same site in the same individual after hours, days or weeks.

In order to better understand microvascular perfusion and LDF signals, and to improve laser Doppler flowmeters, numerical simulations of LDF signals have now become necessary. Numerical simulations of LDF signals can be performed with Monte Carlo methods. The latter rely on computational algorithms using repeated random sampling to compute the simulated results.

2. Monte Carlo simulations

2.1 Introduction

In Monte-Carlo simulations, light transport in tissue is described in the form of separate photons travelling through the sample. On its way, the photon might be scattered at (or in) particles, by which the direction of the photon is changed, or the photon is absorbed. The scattering phenomenon will be determined by suitable angle-dependent scattering functions. When a boundary between two layers, or between the sample and the surrounding medium, or between an internal structure and the surrounding layer, is encountered, the photon might be reflected or refracted. This is determined by the well-known Fresnel relations. In between these events, the photon will propagate, and the optical mean free path in that part of the sample will determine the length of the propagation path. The actual length of the contributions to the path, the angles of scattering, the choice between scattering and absorption, and between reflection and refraction, are determined by random number-based decisions.

Some extra features can be applied to the photons. For instance, photons can be thought of as scattering at particles at rest or at moving particles. This effect will cause a Doppler shift in the frequency of the photons, which can be registered. Afterwards from the Doppler shift distribution of all suitably detected photons the frequency power distribution can be derived. Several models are present for this velocity shift: unidirectional or random flow, various flow profiles and so on. Another option is to use as the light source not a beam impinging from the outside world, but a photon absorption distribution inside the sample. In this way, fluorescence or Raman scattering can be mimicked. When recording the path of

the photons through the sample, one might deduce the path length distribution, and from that the time-of-flight distribution. The latter can be used to predict the distributions of phase delays and modulation depths encountered when performing frequency-modulation experiments.

Further, the distribution of positions where photons were absorbed can be used as the distribution of sources for calculating the photoacoustic response, to be detected using suitable detector elements (or groups of elements, to take interference effects into account) at the surface of the sample.

To start simulating the photon transport, following preparations are needed (see www.demul.net/frits):

- Definition of types of particles (optical properties, concentrations, velocities, etc.)
- Calculation of angle-dependent scattering functions for all types of particles
- Definition of the light source, either a pencil beam or a broad divergent beam or an internal source
- Definition of the sample system, consisting of one or more layers with different contents, with different optical characteristics and velocity profiles; the sample may contain "objects": (arrays of) cylinders, spheres, cones, rectangular blocks, and mirrors. See Figure 2
- Definition of the detection system, consisting of a poly-element detection window, and of its numerical aperture

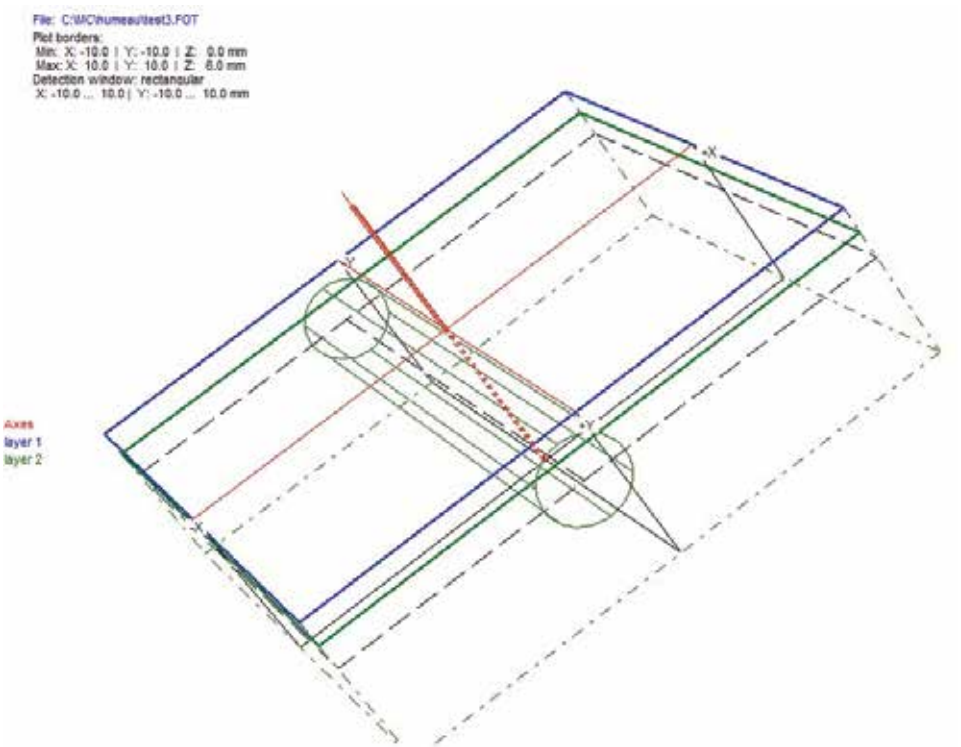


Fig. 2. Structure plot of a two-layer system with a horizontal cylindrical tube, filled with various concentrations of scattering/absorbing particles. Laser light (here pencil beam) injected along Z-axis.

- Definition of the calculation mode: *e.g.* reflection or transmission, or absorption, or a combination of those
- Extra features, to follow the simulations, like LDF, photoacoustics and frequency modulation

These points will be detailed below. A computer package to carry these simulations is available (see the site www.demul.net/frits). All the processes that are presented in this section are dealt with in this computer package. The physical mathematics behind it and detailed explanations can be found in de Mul (1995, 2004) and the site www.demul.net/frits

2.2 Transport algorithms

In order to describe the transport of photons through the sample, one needs algorithms for the various events that the photon may encounter. Those are: scattering or absorption, reflection or refraction at boundaries, and detection. In addition, a mechanism accounting for the destruction of irrelevant photons (*e.g.* photons that have travelled extremely far from the detection window) should be available.

There are two basic algorithms for handling non-zero absorption in layers or particles. Frequently the probability of absorption is taken into account as a "weight factor" for the photon. The cumulative effect of applying these subsequent factors at each scattering event will reduce its overall weight in calculating averages of relevant variables (such as intensity) over a set of emerged photons. An example is the work of Wang & Jacques (1993). An advantage is that no photons will be lost by absorption, which can be of importance when the absorption is relatively strong.

Another algorithm does not make use of weight factors, but applies a "sudden death"-method: the photon is considered to be completely absorbed at once, and will thus be removed from the calculation process. This method might be a bit more time consuming, especially when absorption is not very low in a relative sense, but it offers the advantage to study the positions where the photons actually are absorbed. In this way extra features like photoacoustics or fluorescence response can be studied. In view of this option, we have chosen for the second method (see the site www.demul.net/frits).

2.3 Propagation

The average translation distance L for a photon in a layer or object with scattering particles of varying type, in the case of no absorption, is determined by the inverse of the effective scattering coefficient, which is the sum of the products of the concentrations and scattering cross sections of all types of particles in that layer or object. Now we can deduce the expression for the actual path length Δp:

$$\Delta p = L.\ln(1 - R)\,, \tag{1}$$

where R is a random number ($0 \le R < 1$), used for the probability f_s ($0 < f_s \le 1$) to arrive at a path length Δp:

$$f_s = 1\text{-}\exp(-\Delta p/L). \tag{2}$$

However, this path might end prematurely when a boundary at an interface is met. In this case the path will partially stretch out into the medium at the other side of the interface. When dealing with this part of the path, it should be kept in mind that it has to be corrected

in length according to the mean free path for the photons in the two media. See below for a full account.
Now the probability f_a for absorption by the medium l (layer or block) before the photon has reached the end of path Δp_{eff} ($\leq \Delta p$) can be defined as:

$$f_a = 1 - \exp(\mu_a . \Delta p_{eff}) , \tag{3}$$

where μ_a is the absorption coefficient.
By choosing a fresh random number, the probability for absorption on the effective path can be calculated.
Absorption in the system may have two origins, first taking place within the particles themselves, and secondly the absorption by the medium itself. Together with scattering, this leads to an "average translation length" and an "average absorption length" for the medium.
In a previous paper (Kolinko, 1997) we discussed two equivalent algorithms to determine the remaining path length after crossing an interface.
Figure 3 presents a view of a running simulation in a sample with two layers and two objects.

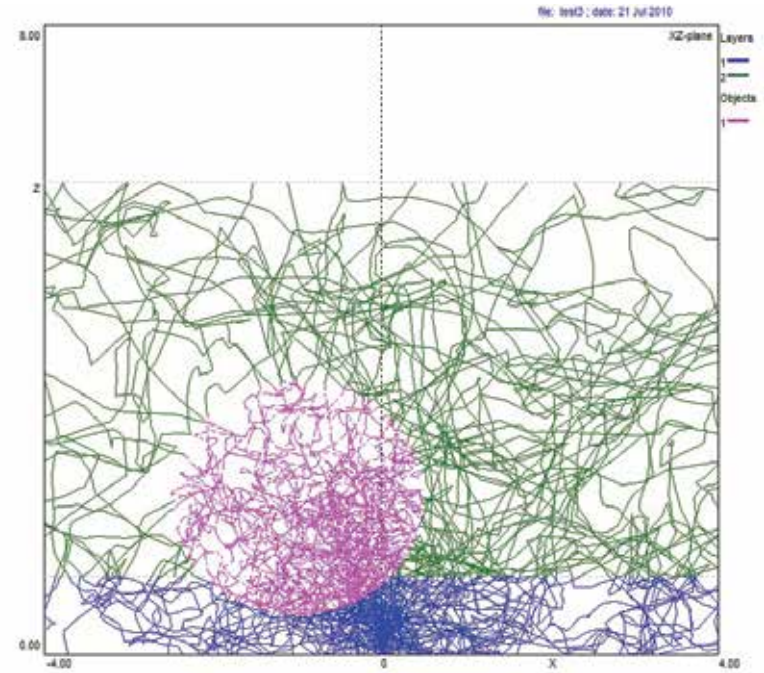


Fig. 3. Running graphics of the simulation process of the structure of Fig. 2. View in YZ-plane. Photons entering around position (0,0,0). The tube (X-direction) and sphere can be seen.

2.4 Scattering

The probability of scattering to the direction given by the angles θ and φ is described by the *scattering function* $p(\theta,\varphi)$. This function is normalized in such a way that the total scattering over the whole 4π solid angle is unity (Figure 4):

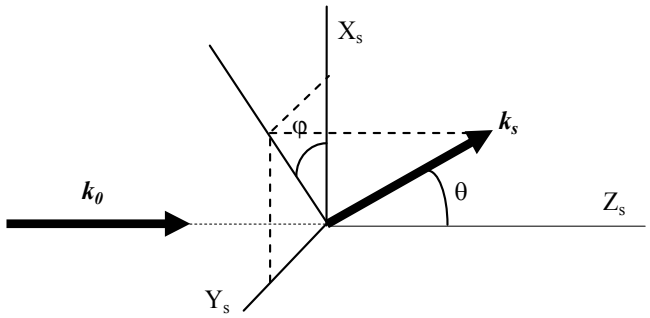


Fig. 4. Basic scattering geometry in the "scattering system" (subscript *s*). The incoming and scattered wavevectors are denoted by $\mathbf{k_0}$ and $\mathbf{k_s}$ respectively. $|\mathbf{k}| = 2\pi/\lambda$, with $\lambda = \lambda_{vacuum} / n$ (n = refractive index of the medium).

$$\int_0^{2\pi} d\varphi \int_0^{\pi} d\theta . p(\theta,\varphi) \sin\theta = 1 \,. \tag{4}$$

For the scattering function, several models are available: Dipole- or Rayleigh-scattering, Rayleigh-Gans scattering, Mie scattering, isotropic or peaked-forward scattering. These scattering functions have been described in many textbooks. We refer here to the standard books of Van de Hulst (1957, 1981). Also models of Henyey and Greenstein (1941) and others are used (see below).

The method of determining the scattering angles θ and φ is:

- For the azimuthal angle φ: $\varphi = R.2\pi$
- For the polar angle θ a normalised cumulative function of the scattering function is used, with values between 0 and 1, and by choosing R, the corresponding θ is calculated

In case polarization effects have to be taken into account, the choice of the angles θ and φ is coupled to the polarization state of the photon.

2.5 Boundaries

Since the program (see www.demul.net/frits) allows for insertion of special structures and objects, like tubes, spheres, mirrors and cones in the layer system, we have to deal with boundaries at flat surfaces (like those between layers) and at curved surfaces.

For flat surfaces, parallel to the layer surface (*i.e.* perpendicular to the Z-axis), the calculation of reflection or refraction angles is according to Snell's law. The fraction of reflected light is given by the Fresnel relations.

For curved surfaces, or flat surfaces not perpendicular to the Z-axis, the construction of local coordinate frames, along the local normal vector, is necessary, which implies foregoing and subsequent coordinate rotations from the laboratory system to the local system and back.

In de Mul (2004, see also www.demul.net/frits) the boundary expressions are derived for following curved surfaces (if applicable, with flat end surfaces at top and bottom):

- Cylinders (parallel to the layer surface, and parallel to the Z-axis, and oblique)
- Arrays of cylinders (linear or two-dimensional)
- Spheres, and two-dimensional arrays of spheres
- Rectangular blocks

- Mirrors (parallel to the layer surface, and parallel to the Z-axis, and oblique)
- (half) toruses

Objects can stretch over layer boundaries. In all cases, carefully the path of a photon has to be followed. Is the photon reflected or refracted, how far does it propagate in the new medium, is there another object within the object, will absorption take place before a scattering event, and (in case of arrays of objects) will the photon propagate from one member of the array set to another?

2.6 Scattering functions

Here we will only mention the expressions for the most commonly used scattering functions. For further study, see Van de Hulst (1957, 1980) or the full report (de Mul, 2004, or website). Important parameters are μ_s, μ'_s, and μ_a, the scattering coefficient, the reduced scattering coefficient and the absorption coefficient, respectively, all expressed in mm^{-1}, with μ_s and μ'_s connected by $\mu'_s = \mu_s\,(1-g)$, g being the averaged scattering polar angle: $g = \langle\cos\theta\rangle$. For tissue, typical μ_s-values are 10-200 mm^{-1} and with typical g-values of 0.90-0.99 this leads to typical μ'_s - values of 1-2 mm^{-1}.

- **Dipolar (Rayleigh)**

With dipolar scattering, the particles are assumed to be so small that light scattered from different oscillating electrical dipoles in the particles will not lead to phase differences upon arrival at the point of detection (Van de Hulst, 1957, 1980). Using standard electromagnetic dipole radiation theory, or a standard Green's functions approach, we may derive for the intensities $I_{//}$ and $I_{\perp}$, proportional to the squares of the field strengths ($I = \frac{1}{2}c\varepsilon_m E^2$), thus:

$$I_{//} = \frac{\alpha^2 k^4}{(4\pi\varepsilon_m)^2 r^2}\cos^2\theta.\cos^2\varphi.\,I_0$$
$$I_{\perp} = \frac{\alpha^2 k^4}{(4\pi\varepsilon_m)^2 r^2}\sin^2\varphi.\,I_0. \tag{5}$$

For natural light the total intensity is given by:

$$I_{nat}(\theta) = \frac{\alpha^2 k^4 I_0}{(4\pi\varepsilon_m)^2 r^2}\,\frac{1+\cos^2\theta}{2}\,. \tag{6}$$

- **Rayleigh-Gans**

When particles grow larger, the phase differences of scattered waves arriving at the detection point from different source points in the scattering medium, cannot be neglected any more. For small differences in the dielectric constant between particles and surrounding medium, the intensity I will be proportional to:

$$I \sim \frac{1+\cos^2\theta}{2}k^4V^2\frac{(m-1)^2}{(2\pi)^2}\left|R(\theta,\varphi)\right|^2, \tag{7}$$

with:

$$R(\theta,\varphi) = \frac{1}{V}\iiint\exp(i\delta).dV\,. \tag{8}$$

The phase-difference δ is given by $\boldsymbol{k} \bullet (\boldsymbol{r}\text{-}\boldsymbol{r}_O)$, where $\boldsymbol{r}$ and $\boldsymbol{r}_O$ are the position vectors from the scattering volume element under consideration and the origin in the sample.

- **Mie**

In principle, the rigorous scattering theory, as developed by Mie (see ref. in Van de Hulst, 1957, 1981), presents analytical expressions for all kind of particles. It departs from the Maxwell equations and solves the scalar part of the wave equation, taking boundary conditions into account. This leads to complicated expressions for the components of Van de Hulst's scattering matrix, which are only tractable when treated numerically. See Figure 5 for an example.

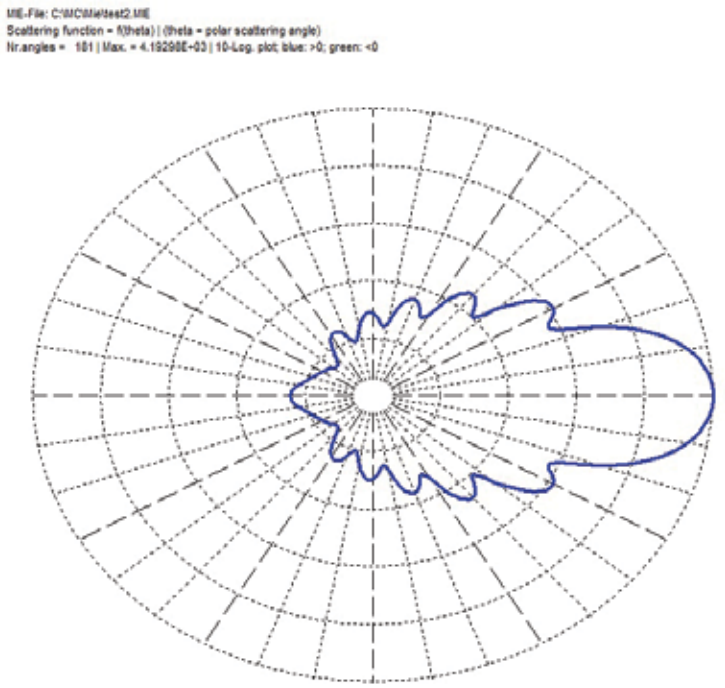


Fig. 5. Example of a MIE-file. Scattering function according to the Mie-formalism (weight 1.0) + Henyey-Greenstein with g=0.80 (weight 0.1).

- **Henyey-Greenstein**

The scattering function of Henyey & Greenstein (1941) originates from the astronomical field, to calculate the scattering by cosmic particle clouds. Since it can be written in a closed analytical form, it can be used as a fast replacement for the Mie-functions. The function reads:

$$p_{HG}(\theta,\varphi) = \frac{1}{4\pi} \frac{1-g^2}{(1+g^2-2g.\cos\theta)^{3/2}}, \tag{9}$$

where g is the averaged cosine of the polar angle θ of the scattering events. This function is normalised to unity upon integration over 4π solid angle. It only describes the angle-dependent behaviour of the scattering. The calculation of the scattering cross section has to be done by other means. One option is to insert the total scattering cross section as obtained

by Mie-scattering (or another approach, if applicable) as a separate factor in the Henyey-Greenstein expression.

- **Other functions**

Other scattering functions are: isotropic scattering, peaked forward, or Gegenbauer (which is an extension of a Henyey-Greenstein-function). We will not deal with those here.

2.7 Light sources

For the injection of photons, one can imagine various mechanisms. Most general is the pencil beam, entering from the top. However, other beam profiles can be used as well. In de Mul (2004, see web site) several options are implemented: pencil beams (perpendicular and oblique), divergent beams, broad parallel beams, ring-shaped beams, isotropic injection and internal point sources (one point or distributed).

Distributed internal sources can be used in simulating Raman or fluorescence scattering, consisting of (1) a simulation of absorptions, and (2) injection of new photons from the positions of absorption.

2.8 Detection

We may distinguish between external detection (at the top or bottom of the sample system: "reflection" or "transmission", or at an internal layer or object boundary) or internal detection (upon an absorption event). In this way, the scattering inside a sphere (a human head?) can be detected.

2.9 Photon path tracking

The tracking of the path of the photon, *i.e.* recording the coordinates of the scattering events and of the intersections with interfaces, can easily result in enormous files. With a scattering coefficient μ_s of about 10-20 mm^{-1} and a g-factor (average of the cosines of the polar scattering angles) of about 0.80 – 0.90, in each mm of the path about $1/\mu_s \approx 10$ scattering events will take place. However, due to the large g-factor, the scattering will be predominantly in forward direction and it will only be after about $1/\mu'_s \approx 1$ mm that the direction of the photon can be considered as randomised. Therefore, in those cases it is better to register only part of the events, namely those at intervals of $1/\mu'_s = 1$ mm, which will decrease the storage space to 144 Mbytes per simulation.

Therefore, the program offers the options of recording the paths at intervals of $1/\mu_s$ or $1/\mu'_s$ (see www.demul.net/frits).

Photons originating from a pencil beam and emerging at equal distances d from the point of injection but at different positions on that ring are equivalent. However, visualisation of those tracks will end up in an un-untwinable bunch. Therefore, to clarify viewing we may rotate the whole paths around the axis of the pencil beam to such an orientation as if the photons all emerged at the same position on the ring, *e.g.* the crossing point with the X-axis.

See Figure 6 for an example of the path tracking method.

2.10 Special Features: laser Doppler flowmetry

Some special features are incorporated in the program (available at www.demul.net/frits). LDF is the oldest feature, built in from the beginning of the development of the program, and meant to support measurements of laser Doppler perfusion flowmetry in tissue.

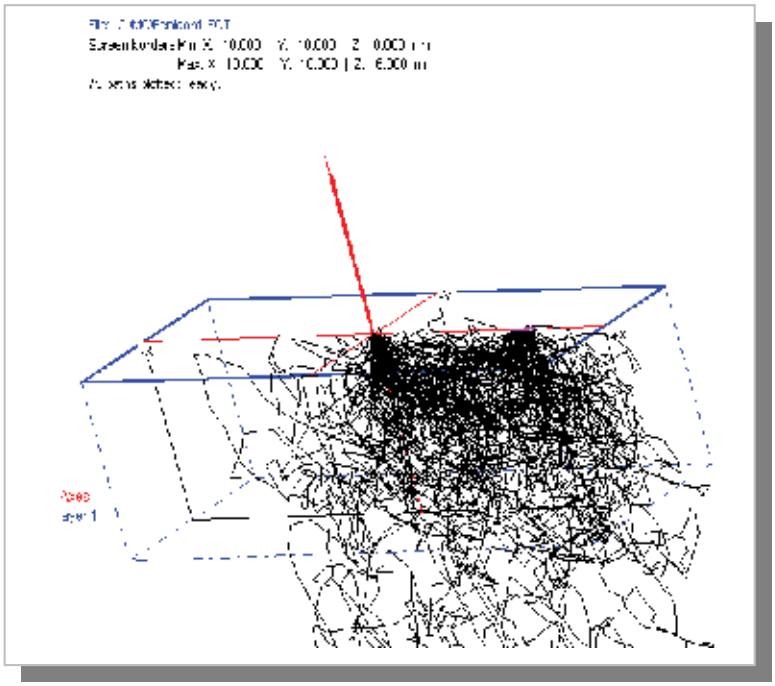


Fig. 6. Photon path tracking: photon "bananas" arising by scattering from beam entrance point to exit area (between 5 and 6 mm). For clarity, all photon paths were rotated afterwards as if the photons had emerged on the +X-axis.

Photoacoustics has been added to simulate the acoustic response to pulsed light. Frequency modulation is a modality adding extra information using path length-dependent phase delay information. Here we only deal shortly with LDF.
As mentioned previously, LDF makes use of the Doppler effect encountered with scattering of photons in particles when those particles are moving. The principles are shown in Figure 7. Using the definitions of the variables given in that Figure, the Doppler frequency is given by:

$$\omega_D = (\mathbf{k}_s - \mathbf{k}_0) \bullet \mathbf{v}\,, \tag{10}$$

and with:

$$|\delta\mathbf{k}| = 2k.\sin\tfrac{1}{2}\theta\,, \tag{11}$$

we find:

$$f_D = \frac{kv}{\pi}\sin\tfrac{1}{2}\theta.\cos\alpha\,. \tag{12}$$

When applied to tissue, frequently the angles θ and α might be considered randomised. This is due to three reasons:

- Preceding scattering by non-moving particles might cause the direction of the photons to be randomised upon encountering moving particles

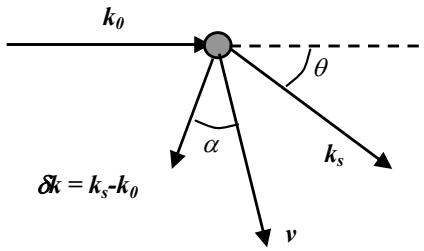


Fig. 7. Principles of LDF. The particle has a velocity **v**. Vectors $\mathbf{k_0}$ and $\mathbf{k_s}$ denote the incoming and scattered light wave vectors, and **δk** is the difference vector.

- Most important moving particles are blood cells in capillaries. Due to the (more or less) random orientation of the capillaries the velocities will have random directions
- Travelling from injection point to detection point, in general the photons will encounter many Doppler scattering effects, with random velocities and orientations

All three effects will broaden the Doppler frequency distribution, which ideally would consist of one single peak, to a smooth distribution as in Figure 8. This means that it is not possible to measure the local velocity, but we only may extract information about the averaged velocity over the measuring volume. The averaging concerns the three effects mentioned above.

There are two options to record these LDF-spectra: homodyne and heterodyne, depending on the relative amount of non-shifted light impinging on the detector. The first is the mutual electronic mixing of the Doppler-shifted signals, and the second is the mixing of those signals including mixing with non-shifted light, which can be overwhelmingly present. The resulting frequency and power spectra (which is the autocorrelation function of the frequency spectrum) will look as sketched in Figure 8.

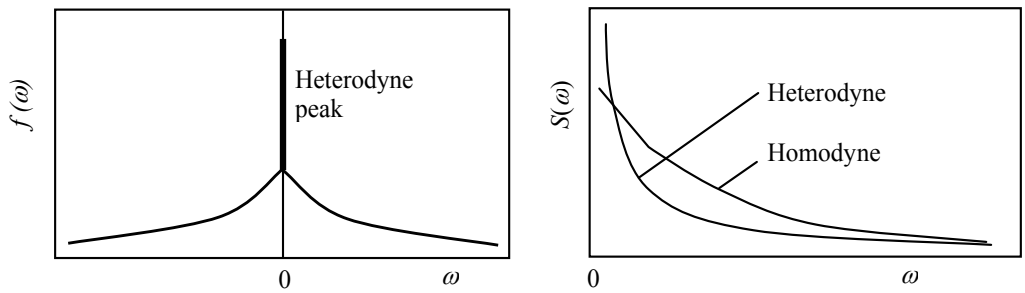


Fig. 8. Homodyne and heterodyne frequency spectra $f(\omega)$ and power spectra $S(\omega)$. Normally the heterodyne peak is much higher than the signals at non-zero frequencies.

To characterize the frequency spectra use is made of moments of the power spectrum $S(\omega)$, defined as:

$$M_n = \int_0^\infty \omega^n . S(\omega) . d\omega \tag{13}$$

and the reduced moments:

$$M_n' = \frac{M_n}{{M_0}^n}. \tag{14}$$

The zeroth moment is the area under the power spectrum itself, and can be considered as proportional to the concentration of moving particles in the measuring volume. Bonner & Nossal (1981) showed that the first moment M_1 is proportional to the averaged flow, while the ratio of the reduced moment $M_1' = M_1 / M_0$ will be proportional to the averaged velocity. Analogously, the reduced moment $M_2' = M_2 / M_0^2$ will be proportional to the average of the velocity-squared.

3. Monte Carlo simulations applied to the microcirculation domain

Monte Carlo simulations have been used since the early nineties to provide a better understanding on LDF measurements. At the beginning, LDF signals were simulated using very simple single layer flow models. Later, complex physical models constructed with the purpose to validate and calibrate laser Doppler flowmeters were proposed. LDF signals were generally simulated in models mimicking the skin and then compared with real data recorded in skin.

Monte Carlo simulations allow studying the influence of tissue parameters (as optical properties and blood velocities) as well as probe configurations and laser light wavelengths on LDF signals. Monte Carlo simulations have already provided information 1) on the blood microcirculatory flow depth measurements in skin (non-invasive measurements) or in other organs (invasive measurements), 2) on the determination of the photon path length, 3) on the dependence of LDF signal on multiple scattering, and 4) on the prediction of the speed distributions of moving particles.

In the LDF domain, Monte Carlo simulations were used, for the first time, by Jentink et al. (1990). It was the first time that the frequency shifts, due to the moving particles, were included in light propagation simulations in scattering and absorbing media. Monte Carlo simulations were used to study the influence of the optical probe configuration and the multiple scattering of photons by moving particles in living tissues. A very simple flow model, consisting of a homogeneous slab with spheres acting as scatterers, has been used. The spheres were moving in random directions, which permitted the simulation of different velocity distributions. The velocity angular distribution was modelled with Mie formula. Light absorption was taken into account by including a constant probability with two terms, one due to the absorption by the spheres, and a smaller one which represented the absorption between spheres. Ring shaped detectors were attached in a concentric arrangement around the light beam and heterodyne detection was assumed. Homodyne detection was ignored. It was observed that, with the increase of source-detector distance, the sampling volume increased whereas the intensity of the signal decreased (Jentink et al., 1990). Therefore, the authors suggested the use of different source-detector separations in order to differentiate perfusion in different skin layers. The limited computational speed available in the early nineties was a serious limitation for the first studies using Monte Carlo simulations.

A more complex skin tissue model, consisting of three different homogeneous layers (epidermis with 0.1 mm thickness, dermis-1 with 0.2 mm thickness, and dermis-2

with 200 mm thickness) was used in Monte Carlo simulations by Koelink et al. (1994). The goal was to determine the sampling volume in LDF measurements using two different wavelengths, 633 nm and 800 nm, and two different source-detector separations, 0.1 and 2 mm. The optical parameters used were 0.015 and 0.01 mm^{-1} for the absorption coefficients for 633 and 800 nm, respectively, for all three layers. The scattering coefficients used were 25 and 15 mm^{-1} for epidermis for 633 and 800 nm, respectively. For dermis-1 and dermis-2, they were of 11.2 and 6.8 mm^{-1} for 633 and 800 nm, respectively. The refractive indexes were 1.5 for epidermis and 1.4 for dermis-1 and for dermis-2. A Henyey-Greenstein phase function with an anisotropy factor g equal to 0.85 for static tissue was used. Microcirculation was assumed in the two deeper layers with a concentration of $1x10^4$ red blood cells/mm^3 in both layers, in random directions. A velocity equal to 1 mm/s was simulated in dermis-1, whereas for dermis-2 the velocity was equal to 1 or 4 mm/s in different simulations. The scattering cross section, σ_s, of the red blood cells used was equal to 25.4 and 15 μm^2 for 633 and 800 nm, respectively. The absorption cross section, σ_a, was equal to 0.065 and 0.042 μm^2 for 633 and 800 nm, respectively. For the scattering phase function of red blood cells the Rayleigh-Gans phase function was applied with an anisotropy factor g equal to 0.985 and 0.98 for 633 and 800 nm, respectively (Koelink et al., 1994). The aim of this work was to distinguish the superficial from the deeper blood vessels microcirculation. Measurements in skin were compared with simulations showing a reasonable agreement (Koelink et al., 1994).
With the purpose to compare real data and Monte Carlo simulations of flow, de Mul et al. (1995) introduced phantoms (physical flow models) in Monte Carlo simulations. The Monte Carlo algorithm used is explained in Section 2. Unlike biological tissues, a phantom permits the control on LDF measurements of relevant parameters, such as scatterers velocity and concentration, optical properties, and so on. Two models were studied: a liquid flow model consisting of a set of liquid layers, and a solid model based on gelatine layers, both with a concentration of 1.25 to 1.9 x 10^6 mm^{-3} of polystyrene spheres acting as scatterers. The goal was also to investigate the relationship between source-detector separation and the photons sampling depth, for different incident angles of the laser beam. The angular scattering distribution for the polystyrene spheres was given by the Mie formulas with an anisotropy factor g = 0.91. The scattering and absorption cross section of the spheres were set to 5.5 and 0.03 μm^2, respectively, at 780 nm. The effects of homodyne and heterodyne scattering were also investigated. The comparisons showed reasonable to good agreement between simulations and real measurements using the phantom (de Mul et al., 1995).
A more complex fluid model applied to Monte Carlo simulations was presented by Steenbergen & de Mul (1997). In order to study the phantom capability to mimic real tissue, the model had optical properties and layered structure similar to those of living tissues. The Monte Carlo algorithm used is explained in Section 2. This phantom, consisting of scattering and absorbing films separated by matching oil, was used to evaluate the consequences of the stratified structure of the phantom in comparison with real tissues, that have no mismatching layers. The phantom was modelled as a semi-infinite repetitive laminate of scattering and absorbing layers with 0.08 mm thickness separated by a transparent colourless layer of resin or oil with 0.005 mm. The laser light source entered the phantom as a pencil beam and the scattered light was detected in a concentric region (radius equal to 2.5 mm) with the pencil beam. Concerning the optical properties, the angular scattering function used was the Henyey-Greenstein distribution with an anisotropy factor g = 0.9. The

scattering coefficient used was μ_s = 5, 10, 20 and 40 mm^{-1} which gave a reduced scattering coefficient equal to μ'_s = 0.5, 1, 2 and 4 mm^{-1} respectively. The absorption coefficient used was μ_a = 0.005 mm^{-1}. The refractive index of the phantom matrix was 1.52. The algorithm used was similar to the one developed by de Mul et al. (1995). When the photon travelled between a transparent layer and a scattering layer, a new path length was determined using a random generator (the path length which has been interrupted by the medium boundary was disregarded). At the end, the mismatching problems revealed not to being as restrictive as it might have been expected (Steenbergen & de Mul, 1997).

The coherence effects in the detection of Doppler signals have also been investigated with Monte Carlo simulations. Measurements in gelatin phantoms with polystyrene scatterers was built to mimic skin tissue characteristics and for calibration and standardization of perfusion tools (de Mul et al., 1999). Homodyne and heterodyne detection were investigated. For the homodyne experiments, a phantom consisting of just one moving layer with 11 mm thickness was used. For the heterodyne experiments, the phantom consisted of one static and one moving layer with 4 and 11 mm thickness, respectively, with the same scatterers concentration. The simulation model consisted of five layers: an air gap between probe and first layer, a fixed layer, another air gap between the first and second layer, a moving layer, and an absorbing layer. The homodyne simulations did not match the corresponding measurements quite properly, showing a too broad Doppler power spectrum that was, in part, reduced by the implemented coherence correction (de Mul et al., 1999).

The influence of the optical properties and the source-detector separation on the sampling depth in LDF measurements were also studied using a sophisticated tissue-like phantom (Larsson et al., 2002). Monte Carlo simulations mimicking the phantom were used to validate the measurements. The tissue phantom and the simulation model consisted of a set of parallel static layers (95 μm thickness) with different optical properties, separated by rotatable moving layers (20 to 22 μm thickness). A laser source of 632.8 nm was used and the backscattered light was guided to the detector by 7 fibres, with a numerical aperture of 0.37 and a diameter of 230 μm, arranged in a row. Hollow polystyrene microspheres, with a diameter of 1 μm, were used as scatterers, and optical absorbers were added to the static layers. The phantom had 4 windows, where the probe could be placed, each one with different combinations of optical properties (14.66 and 0.212 mm^{-1}, 44.85 and 0.226 mm^{-1}, 14.8 and 0.0405 mm^{-1} and 45.55 and 0.0532 mm^{-1} for μ_a and μ_s, respectively). The probability distribution of the photon scattering angle was modelled using the Henyey-Greenstein phase function with an anisotropy factor g=0.815. Two models were simulated: 1) with one single rotating disk having 14.66 and 0.212 mm^{-1} for μ_a and μ_s, respectively, for two different velocities (0.7 and 2.2 mm/s); and 2) with the same optical properties as the four windows described above, where single moving disks and multiple moving disks at 1 mm/s could be modelled. Measured and simulated data showed good correlation and the simulations showed that the sampling depth decreases with the increase of μ_a or μ_s and it increases with the increase of source-detector fibre separation (Larsson et al., 2002).

Other Monte Carlo simulations were used to predict sampling depth of light scattering in skin, in different situations. The sampling depth for a laser Doppler perfusion imaging system, using a simplified model with a single moving layer, was also analyzed (Rajan et al., 2008).

Other authors investigated the use of a high power laser source (20 mW), with 785 nm laser light source and a source-detector distance of 4 mm (Clough et al., 2009).
Based on Monte Carlo simulations of light propagation, the measurement depth/volume was also estimated for various tissues models (muscle, liver, gray matter, white matter, and skin). Both probe-based and imaging systems were studied, at the wavelengths of 453 nm, 633 nm and 780 nm (Fredriksson et al., 2009). Thus, using Monte Carlo simulations, typical measurement depths and volumes for simulated perfusion have been presented, for various types of biological tissues and system setups. The simulations were not compared with *in vivo* measurements (Fredriksson et al., 2009).
A skin model, that can be used to estimate skin-like tissue perfusion in absolute units, was also presented (Fredriksson et al., 2008). The impact on LDF measurements of parameters such as layers thickness, blood concentrations, melanin concentration in epidermis, μ_a, μ_s and g for blood and skin layers were evaluated with Monte Carlo simulations. The simulated spectra generated for 7000 different skin configurations, for two different source-detector separations (0.25 and 1.2 mm), and a 780 nm laser light source, were compared with *in vivo* data. The goal was to validate the best fit model with the measured spectra. In this study, the skin model had 6 layers (epidermis, papillary dermis, superior blood net, reticular dermis, inferior blood net, and subcutis). Moreover, different blood concentrations were chosen with three representative velocities (0.3, 3.0, and 30 mm/s) in each layer, and with a parabolic profile (between 0 and twice the mean velocity of the scattering blood component). The simulated source and detector were configured to mimic the probe used. The calculated spectra were compared with measurements carried out on the forearm and on the finger pulp skin, without heating, and with heat provocations on the forearm skin. At each scattering occasion it was decided, based on the concentration of blood, the scattering coefficient of the blood and the static matrix, if the photon was to be scattered by the static matrix, or by a moving red blood cell, causing Doppler shift. The velocity component (0.3, 3, or 30 mm/s) that produced the Doppler shift, in simulations, was randomly chosen based on the scatterers concentration. The optical power spectrum for the two source-detector separations was calculated. The distribution of accumulated frequency shifts for the detected photons and the Doppler power spectrum was also calculated as the autocorrelation of the optical power spectrum. For the photon launch, a variance reduction method, the implicitly capture, was used. In this method, after the first scattering, the photon was splitted into 50 new photons, each one with 1/50 of the weight of the original photon and it was scattered in random directions. Then, in the following successive Doppler scattering occasions, the photon was splitted again into two new photons, if the total frequency shift of the photon exceeded nx10/6 kHz (n=1, 2, …, 6). One photon could thus be splitted into, at most, $50x2^6 = 3200$ photons. A good combination of thickness of the model and of blood concentrations that produced simulated Doppler power spectra that agreed well with measured Doppler power spectra, was found (Fredriksson et al., 2008).
Methods for photon path length determination in LDF have also been proposed by Monte Carlo simulations (Jakobsson & Nilsson, 1993; Nilsson et al., 2002; Larsson et al., 2003; Varghese et al., 2007). Thus, Jakobsson & Nilsson (1993) used one-layer models of skin, liver and brain tissues to study path length distributions for different probe geometries at a wavelength of 633 nm. The 3D pathways of single photons were computed and stored. Information as the penetration depth, sampling depth, and total photon path length

between source and detector were stored. The Henyey-Greenstein phase function was used as the density function of the scattering angle. The tissues optical properties μ_a, μ_s and g (in cm^{-1} for μ_a and μ_s) were respectively set equal to 2, 188 and 0.8 for skin, 2.3, 313 and 0.68 for liver, 1.3, 48.8 and 0.96 for brain. For the whole blood (unmodified blood), the values were respectively equal to 18, 320 and 0.99. For each distance travelled by a photon, the probability of absorption was calculated. If absorption occurred, the photon pathway was terminated. Different perfusion profiles were simulated in skin tissues model and the optical parameters (μ_a, μ_s, g) were changed stepwise, one by one. The influence of different red blood cell concentrations, in skin tissue was also simulated.

Using a one-layer model with a wide range of optical properties, relevant to human skin, the average path length for various source detector separations up to 2 mm was simulated, using Monte Carlo methods (Nilsson et al., 2002). The Monte Carlo simulation software used was developed by de Mul (de Mul et al., 1995). A reference space model built to develop path length estimation methods with 144 different sets of optical properties (μ_a, μ_s, and g) and a validation space model for evaluation of the accuracy of the path length estimations methods with 75 different sets of optical properties were defined. Different methods for predicting the path length were investigated. A multiple polynomial regression method, based on spatially resolved diffuse reflectance, proved to be the most effective in predicting the average path length as a function of source-detector separation (Nilsson et al., 2002).

Larsson et al. (2003) estimated the photon path length and optical properties for source detector separation up to 1.61 mm. They used a simulated model consisting of a semi-infinite slab, with 100 mm thickness, with a low concentration of moving scatterers and with a constant velocity (v = 1 mm/s). de Mul software (de Mul et al., 1995) was used and the scattering events were simulated with the modified Henyey-Greenstein phase function for a 632.8 nm laser source. Four different sets of optical properties were simulated. Measurements, *in vivo*, at different human skin sites and, *in vitro*, in two phantoms were used together with simulations in order to obtain the estimated μ_a, the estimated reduced scattering coefficient μ'_s, estimated photon path length and normalized and linearized LDF perfusion. The path length estimations were applied for normalization of the estimated perfusion, removing its optical properties dependency (Larsson et al., 2003).

The optical path lengths of shifted and unshifted light, as well as the path length dependent Doppler broadening were measured in a two-layer tissue phantom (with a superficial static layer of different thickness) with a phase modulated low coherence Mach-Zehnder interferometer (Varghese et al., 2007). Validation of the experimentally determined thickness of the static layer and the optical path length distributions were done with the Doppler Monte Carlo methods. The simulated phantom consisted of a static scattering layer with variable thickness (between 0.1 and 0.9 mm), between two static glass layers with 0.15 mm, a dynamic layer with 20 mm, and a fifth layer with high absorption characteristics (μ_a = 10 mm^{-1}). The static and dynamic scattering layers had the same optical properties as they had the same scatterers - polystyrene sphere suspension. The refractive index was 1.33, μ_a and μ'_s were set to 0.001 and 2 mm^{-1}, respectively, and g = 0.85 for Henyey-Greenstein scattering phase function. The thickness of the static layer was estimated from the minimum optical path length of Doppler-shifted light. A good agreement between experimentally

determined thickness of the static layer and Monte Carlo simulation was obtained (Varghese et al., 2007). The method was able to measure path length resolved information of non-shifted and shifted Doppler fractions of photons (Varghese et al., 2007).

Using Monte Carlo simulations, a few authors worked on the prediction of the speed distributions of moving particles based on the laser Doppler spectrum decomposition (Fredriksson et al., 2006; Larsson & Stromberg, 2006; Liebert et al., 2006). As the velocity depends on the dimension of the blood vessels, the prediction of the speed distributions could lead, *in vivo*, to differentiate between capillary and arterial blood flow. Various uniform and Gaussian speed distributions of particles, moving in the turbid media, were simulated in order to study the relation between the calculated speed distributions of moving particles and the simulated distribution using Monte Carlo simulations (Liebert et al., 2006). The Henyey-Greenstein phase function was used for the calculations and several anisotropy factors of the medium were simulated (Liebert et al., 2006). The optical properties, μ_a, μ'_s and refractive index n were set to 0.01 mm^{-1}, 1 mm^{-1} and 1.4, respectively. A laser light source of 780 nm was used. The backscattered photons were collected in concentric ringshaped detector at 1 mm from the source and a concentration of 1% of moving scatterers was used for simulations. The theoretical background of single and multi-scattering were presented, but it was validated by Monte Carlo simulations for single-scattering, only. The calculated speed distribution of moving particles matched with the assumed simulated Gaussian distributions of the moving particles speeds (Liebert et al., 2006).

An algorithm for velocity resolved perfusion measurements, that characterizes the microvascular blood flow in three different velocities, was suggested by Larsson & Strömberg (2006). This algorithm was derived by fitting a set of predefined Monte Carlo simulated, single velocity spectra, to a measured, multiple velocity LDF spectrum, based on single Doppler scattering event. The proposed method yields three concentration measurements, each associated with a predefined, physiologically relevant, absolute velocity. A perfusion phantom with a microsphere solution (or diluted blood), and with single or double-tube flow was used for validation. A parabolic flow profile, with average flow velocity equal to the real flow velocity, was simulated using the two phantoms with different experimental setups. Mie theory was used for the microspheres scattering angle distribution and Gegenbauer kernel phase function was used for blood. This study showed that the LDF signal can be separated into, at least, three different velocity regions (Larsson & Strömberg, 2006).

Another method based on Doppler power spectrum decomposition into a number of flow velocity components, measured in absolute units (mm/s), was proposed (Fredriksson et al., 2006). With Monte Carlo simulations, the number of Doppler shifts, and the total Doppler shift, were recorded for each detected photon. From the simulated optical Doppler spectrum resulting from the detected photons, the Doppler power spectrum was calculated. The shift distributions thus obtained were used for measured, and calculated, spectra comparisons. With the simulated shift distribution, a Doppler power spectrum was calculated, originating from a certain combination of velocity components using a mathematical model. The non linear Levenberg-Marquardt optimization method was used to fit the calculated and the measured Doppler power spectra, giving the set of velocity components in the measured sample. Evaluation of the method was achieved with a multi-tube flow phantom, perfused

with polystyrene microspheres infusion, or human blood. The results showed good velocity components estimations for low velocities and low concentrations of moving scatterers, but the opposite was found for high velocities and high concentrations. The reason for that was attributed to the physical characteristics of the phantom used (Fredriksson et al., 2006).

A recent study (Binzoni et al., 2009) suggested a reinterpretation of the Monte Carlo simulation in order to obtain more information, relevant to LDF measurements. The authors suggested a method for photo-electric current determination in Monte Carlo simulations which will allow that any algorithm used in real LDF instrument could be tested and validated.

Our research group is using Monte Carlo simulations in order to validate a new laser Doppler flowmeter prototype for depth flow discrimination in skin. This prototype uses different wavelengths (635, 785, and 830 nm) and different source-detector fibre distances (0, 0.14, 0.25, and 1.2 mm). The prototype is evaluated, *in vitro*, in a phantom consisting of six layers of Teflon® microtubes (with 0.3 and 0.76 mm inner and outer diameter, respectively). Skimmed milk is used as the moving fluid (Figueiras et al., 2010). Milk has been chosen because it has various components that act as scatterers (carbohydrates, fat, and protein). Moreover, it does not sediment like microspheres, and it has similar behaviour to intralipid solutions (Waterworth et al., 1995). Finally, milk is easier for handling than blood and, besides, it is cheaper. However, as milk is unstable, we use the same solution of milk for one day only. Milk is pumped in the microtubes with a motorized syringe with different velocities: 1.56, 3.12, 4.68, 6.25, 7.78, and 9.35 mm/s. The prototype will also be evaluated, *in vivo*, in healthy human subjects. Real measurements and simulations will then be compared.

For the three wavelengths and for the four different source-detector fibre separations, LDF simulations are carried out with a skin model similar to the one proposed by Fredriksson et al. (2009), and with the phantom. For the skin layers, μ_a is set equal to 0.15, 0.1, and 0.0122 mm^{-1}, for 635, 785, and 830 nm, respectively. μ_s is set equal to 20, 13, and 18 mm^{-1} for 635, 785, and 830 nm, respectively. Regarding the blood optical properties, μ_a is set equal to 0.34, 0.5, and 0.52 mm^{-1}, for 635, 785, and 830 nm, respectively, and μ_s is set equal to 16, 13, and 11 mm^{-1} for 635, 785, and 830 nm, respectively. Concerning the phantom simulations with a 635 nm laser source, the refractive index for milk is 1.346, μ_a and μ_s are 0.00052 and 52 mm^{-1}, respectively. For the tubes, the refractive index is 1.367, μ_a and μ_s are equal to 0.001 and 167 mm^{-1}, respectively. The software used was developed by de Mul (1995). The mean photon Doppler shifted depths obtained for the modelled skin and for the phantom are presented in Figure 9. For skin simulations, it can be seen that the mean Doppler-shifted photon depth increases with the fibre source-detector separation and with the laser wavelength. The results obtained for skin are similar to the ones obtained by Fredriksson et al. (2009). The results obtained for the phantom are higher when compared with skin simulated results. We can explain these differences by the physical structure of the phantom that is different from the physical structure of the skin. Moreover, the phantom optical properties are different from the ones of simulated skin. Figure 10 shows a software window with the power spectrum obtained in the phantom for a wavelength of 635 nm and for different velocities: 7.78 mm/s (light green); 6.25 mm/s (red), 4.68 mm/s (pink); 3.12 mm/s (green) and 1.56 mm/s (blue). The spectra were taken with a 1.2 mm source detector separation. As it was expected the power spectrum increases with the velocity.

The mentioned works prove the great importance of Monte Carlo simulations applied to LDF measurements.

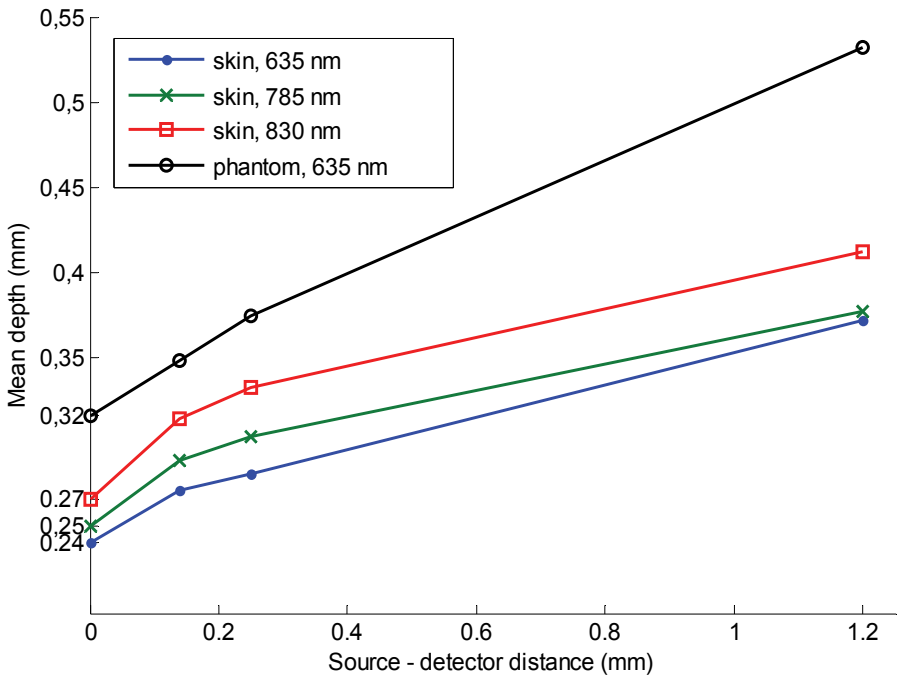


Fig. 9. Mean photon Doppler shifted depths obtained in skin simulations for different wavelengths: 635 nm (blue curve) with μ_a and μ_s equal to 0.15 and 20 mm^{-1}, respectively for skin layers, and with μ_a and μ_s equal to 0.34 and 16 mm^{-1}, respectively for blood; 785 nm (green curve) with μ_a and μ_s equal to 0.1 and 13 mm^{-1}, respectively for skin layers, and with μ_a and μ_s equal to 0.5 and 13 mm^{-1}, respectively for blood; and 830 nm (red curve) with μ_a and μ_s equal to 0.0122 and 18 mm^{-1}, respectively for skin layers, and with μ_a and μ_s equal to 0.52 and 11 mm^{-1}, respectively for blood. Phantom simulations are also shown for 635 nm with milk pumped at 1.56 mm/s with μ_a and μ_s equal to 0.00052 and 52 mm^{-1}, respectively for milk, and equal to 0.001 and 167 mm^{-1}, respectively for the tubes. Different source-detector separations (0, 0.14, 0.25 and 1.2 mm) are used.

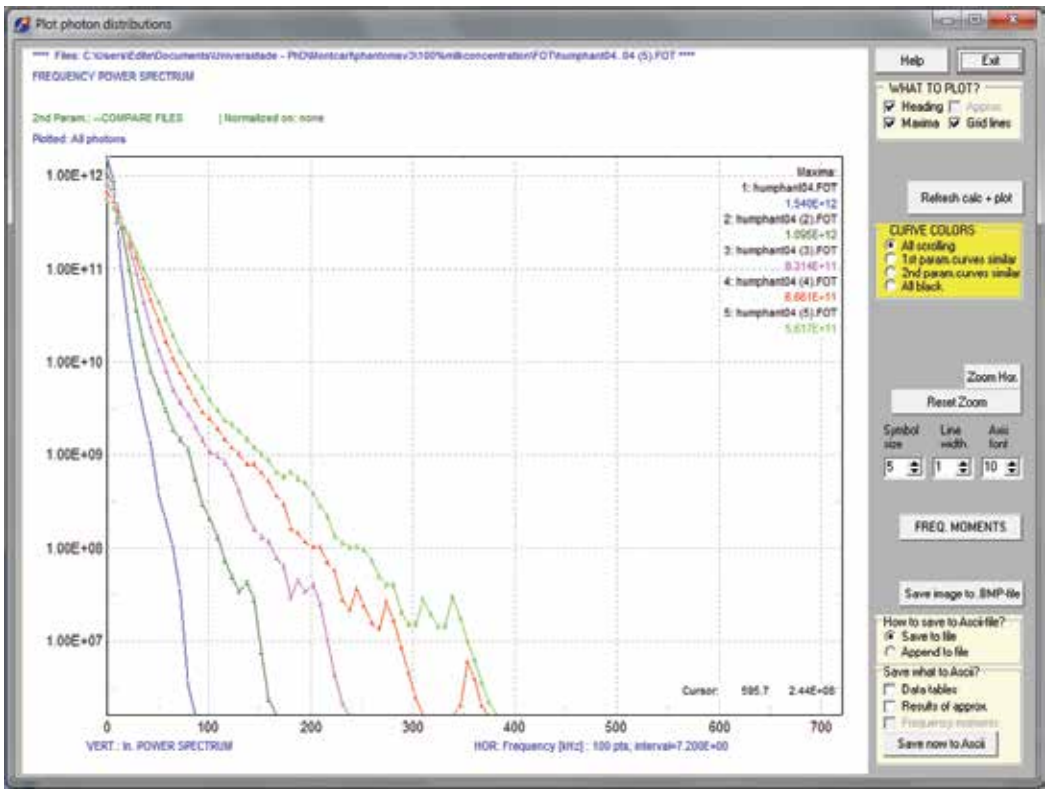


Fig. 10. Software window (software from de Mul, see www.demul.net/frits) showing the power spectrum obtained with a 635 nm laser source, in the phantom, for different velocities (light green: 7.78 mm/s; red: 6.25 mm/s, pink: 4. 68 mm/s, green: 3.12 mm/s and blue: 1.56 mm/s) and for 1.2 mm source-detector separation.

5. References

Binzoni, T., Leung, T. S. & Van De Ville, D. (2009). The photo-electric current in laser-Doppler flowmetry by Monte Carlo simulations. *Physics in Medicine and Biology*, Vol. 54, No. 14, N303-318.

Bonner, R. & Nossal, R. (1981). Model for laser Doppler measurements of blood flow in tissue. *Applied Optics*, Vol. 20, No. 12, 2097-2107.

Cal, K., Zakowiecki, D. & Stefanovska, J. (2010). Advanced tools for *in vivo* skin analysis. *International Journal of Dermatology*, Vol. 49, No. 5, 492-499.

Clough, G., Chipperfield, A., Byrne, C., de Mul, F. & Gush, R. (2009). Evaluation of a new high power, wide separation laser Doppler probe: potential measurement of deeper tissue blood flow. *Microvascular Research*, Vol. 78, No. 2, 155-161.

de Mul, F. F. M., Koelink, M. H., Kok, M. L., Harmsma, P. J., Graaf, R. & Aarnoudse, J. G. (1995). Laser Doppler velocimetry and Monte Carlo simulations on models for blood perfusion in tissue. *Applied Optics*, Vol. 34, No. 28, 6595-6611.

de Mul, F. F. M., Steenbergen, W. & Greve, J. (1999). Doppler Monte Carlo simulations of light scattering in tissue to support laser Doppler perfusion measurements. *Technology and Health Care*, Vol. 7, No. 2-3, 171-183.

de Mul, F. F. M. (2004). Monte-Carlo simulations of light transport in turbid media, In: *Handbook of Coherent Domain Optical Methods, Biomedical Diagnostics, Environment and Material Science*, Tuchin, Valery V. (Ed.), 465–533, Kluwer Publishers, ISBN 1402075766, Dordrecht, the Netherlands.

Figueiras, E., Requicha Ferreira, L.F., Humeau, A. (2010). "Phantom validation for depth assessment in laser Doppler flowmetry technique", Proceedings of EOS - Topical Meeting on Diffractive Optics, paper 2413, Koli (Finland), 14-18 February 2010; proc. ISBN: 978-3-00-024193-2.

Fredriksson, I., Larsson, M. & Strömberg, T. (2006). Absolute flow velocity components in laser Doppler flowmetry, *Proceedings of SPIE 6094 – Optical Diagnostics and Sensing IV*, paper 60940A, 48-59, USA, January 2006, SPIE, San Jose.

Fredriksson, I., Larsson, M. & Strömberg, T. (2008). Optical microcirculatory skin model: assessed by Monte Carlo simulations paired with *in vivo* laser Doppler flowmetry. *Journal of Biomedical Optics*, Vol. 13, No. 1, 014015.

Fredriksson, I., Larsson, M. & Strömberg, T. (2009). Measurement depth and volume in laser Doppler flowmetry. *Microvascular Research*, Vol. 78, No. 1, 4-13.

Henyey, L.G. & Greenstein, J. L. (1941). Diffuse radiation in the galaxy. *The Astrophysical Journal*, Vol. 93, 70-83.

Humeau, A., Koïtka, A., Abraham, P., Saumet, J. L., & L'Huillier, J. P. (2004). Spectral components of laser Doppler flowmetry signals recorded in healthy and type 1 diabetic subjects at rest and during a local and progressive cutaneous pressure application: scalogram analyses. *Physics in Medicine and Biology*, Vol. 49, No. 17, 3957-3970.

Humeau, A., Steenbergen, W., Nilsson, H. & Strömberg, T. (2007). Laser Doppler perfusion monitoring and imaging: novel approaches. *Medical and Biological Engineering and Computing*, Vol. 45, No. 5, 421-435.

Jakobsson, A. & Nilsson, G. E. (1993). Prediction of sampling depth and photon pathlength in laser Doppler flowmetry. *Medical and Biological Engineering and Computing*, Vol. 31, No. 3, 301-307.

Jentink, W., de Mul, F. F. M., Hermsen, R. G. A. M., Graaf, R. & Greve, J. (1990). Monte Carlo simulations of laser Doppler blood flow measurements in tissue. *Applied Optics*, Vol. 29, No. 16, 2371-2381.

Koelink, M. H., de Mul, F. F. M., Greve, J., Graaf, R., Dassel, A. C. M. & Aarnoudse, J. G. (1994). Laser Doppler blood flowmetry using two wavelengths: Monte Carlo simulations and measurements. *Applied Optics*, Vol. 33, No. 16, 3549-3558.

Kolinko, V. G., de Mul, F. F. M., Greve J. & Priezzhev A. V. (1997). On refraction in Monte-Carlo simulations of light transport through biological tissues. *Medical and Biological Engineering and Computing*, Vol. 35, No. 3, 287-288.

Larsson, M., Steenbergen, W. & Strömberg, T. (2002). Influence of optical properties and fiber separation on laser Doppler flowmetry. *Journal of Biomedical Optics*, Vol. 7, No. 2, 236-243.

Larsson, M., Nilsson, H. & Strömberg, T. (2003). *In vivo* determination of local skin optical properties and photon path length by use of spatially resolved diffuse reflectance with applications in laser Doppler flowmetry. *Applied Optics*, Vol. 42, No. 1, 124-134.

Larsson, M. & Strömberg, T. (2006). Towards a velocity-resolved microvascular blood flow measure by decomposition of the laser Doppler spectrum. *Journal of Biomedical Optics*, Vol. 11, No. 1, 14024–14033.

Liebert, A., Zolek, N. & Maniewski, R. (2006). Decomposition of a laser-Doppler spectrum for estimation of speed distribution of particles moving in an optically turbid medium: Monte Carlo validation study. *Physics in Medicine and Biology*, Vol. 51, No. 22, 5737-5751.

Merz, K. M., Pfau, M., Blumenstock, G., Tenenhaus, M., Schaller, H. E., Rennekampff, H. O. (2010). Cutaneous microcirculatory assessment of the burn wound is associated with depth of injury and predicts healing time. *Burns*, Vol. 36, No. 4, 477-482.

Nilsson, H., Larsson, M., Nilsson, G. E. & Strömberg, T. (2002). Photon pathlength determination based on spatially resolved diffuse reflectance. *Journal of Biomedical Optics*, Vol. 7, No. 3, 478-485.

Öberg, P. A. (1990). Laser-Doppler flowmetry. *Critical Reviews in Biomedical Engineering*, Vol. 18, No. 2, 125-163.

Rajan, V., Varghese, B., van Leeuwen, T. G. & Steenbergen, W. (2008). Influence of tissue optical properties on laser Doppler perfusion imaging, accounting for photon penetration depth and the laser speckle phenomenon. *Journal of Biomedical Optics*, Vol. 13, No. 2, 024001.

Rajan, V., Varghese, B., van Leeuwen, T. G. & Steenbergen, W. (2009). Review of methodological developments in laser Doppler flowmetry. *Lasers in Medical Science*, Vol. 24, No. 2, 269-283.

Ray, S. A., Buckenham, T. M., Belli, A. M., Taylor, R. S. & Dormandy, J. A. (1999). The association between laser Doppler reactive hyperaemia curves and the distribution

of peripheral arterial disease. *European Journal of Vascular and Endovascular Surgery*, Vol. 17, No. 3, 245-248.

Smit, J. M., Zeebregts, C. J., Acosta, R. & Werker, P. M. (2010). Advancements in free flap monitoring in the last decade: a critical review. *Plastic and Reconstructive Surgery*, Vol. 125, No. 1, 177-185.

Steenbergen, W. & de Mul, F. (1997). New optical tissue phantom, and its use for studying laser Doppler blood flowmetry, *Proceedings of SPIE – Optical and Imaging Techniques for Biomonitoring III*, Vol. 3196, ISBN: 9780819466440, Italy, September 1997, SPIE, Italy.

Van de Hulst, H. C., Light Scattering by Small Particles, Dover Publications, New York, USA, 1957, 1981, ISBN 0-486-64228-3.

Varghese, B., Rajan, V., van Leeuwen, T. G. & Steenbergen, W. (2007). Discrimination between Doppler-shifted and non-shifted light in coherence domain path length resolved measurements of multiply scattered light. *Optics Express*, Vol. 15, No. 20, 13340-13350.

Wang, L. and Jacques, S. L. (1993). Hybrid model of Monte-Carlo simulation and diffusion theory for light reflectance by turbid media. *Journal of the Optical Society of America A*, Vol. 10, No. 8, 1746-1752.

Waterworth, M. D., Tarte, B. J., Joblin, A. J., van Doorn, T. & Niesler, H. E. (1995). Optical transmission properties of homogenized milk used as a phantom material in visible wavelength imaging. *Australasian Physical and Engineering Sciences in Medicine*, Vol. 18, No. 1, 39-44.

Yamamoto-Suganuma, R. & Aso, Y. (2009). Relationship between post-occlusive forearm skin reactive hyperaemia and vascular disease in patients with Type 2 diabetes--a novel index for detecting micro- and macrovascular dysfunction using laser Doppler flowmetry. *Diabetic Medicine*, Vol. 26, No. 1, 83-88.

Ziegler, S., Gschwandtner, M., Zöch, C., Barth, A., Minar, E., Rüdiger, H. & Osterode, W. (2004). Laser Doppler anemometry distinguishes primary Raynaud phenomenon from VWF syndrome. *Microvascular Research*, Vol. 68, No. 3, 203-208.

Part 2

Electromagnetics

8

Generation and Resonance Scattering of Waves on Cubically Polarisable Layered Structures

Lutz Angermann[1] and Vasyl V. Yatsyk[2]
[1]*Institut für Mathematik, Technische Universität Clausthal, Erzstraße 1, D-38678 Clausthal-Zellerfeld*
[2]*Usikov Institute of Radiophysics and Electronics, National Academy of Sciences of Ukraine, Kharkov*
[1]*Germany*
[2]*Ukraine*

1. Introduction

In this paper, mathematical models for the analysis of processes of generation and resonance scattering of wave packets on a transversely inhomogeneous, isotropic, cubically polarisable, non-magnetic, linearly polarised (E polarisation) medium with a non-linear, layered dielectric structure, and methods of their numerical simulation are considered. In general, electromagnetic waves in a non-linear medium with a cubic polarisability can be described by an infinite system of non-linear differential equations. In the study of particular non-linear effects it proves to be possible to restrict the examination to a finite number of equations, and also to leave certain terms in the representation of the polarisation coefficients, which characterise the physical problem under investigation.
Here we investigate the situation where the incident field consists of a packet of three waves oscillating with single, double and triple frequency. An intense field at the basic frequency leads to the generation of the third harmonic, i.e. of a field at the triple frequency. In this case it is possible to reduce the mathematical model to a system of two equations, where only the non-trivial terms in the expansion of the polarisation coefficients are taken into account (see Angermann & Yatsyk (2010)). The consideration of a weak field at the double frequency or at both the double and triple frequencies allows to analyse its influence on the generation process of the third harmonic. In this situation, the mathematical model consists of three differential equations.
The rigorous formulation finally leads to a system of boundary-value problems of Sturm-Liouville type, which can be equivalently transformed into a system of one-dimensional non-linear integral equations (defined along the height of the structure) with respect to the complex Fourier amplitudes of the scattered fields in the non-linear layer at the basic and multiple frequencies. In the paper both the variational approach to the approximate solution of the system of non-linear boundary-value problems of Sturm-Liouville type (based on the application of a finite element method) and an iterative scheme of the solution of the system of non-linear integral equations (based on the application of a quadrature rule to each of the non-linear integral equations) are considered.

The numerical simulation of the generation of the third harmonic and the resonance scattering problem by excitation by a plane wave packet passing a non-linear three-layered structure is described. Results of the numerical experiments for the values of the non-linear dielectric constants depending on the given amplitudes and angles of the incident fields are presented. Also the obtained diffraction characteristics of the scattered and generated fields are discussed. The dependence characterising the portion of generated energy in the third harmonic on the values of the amplitudes of the excitation fields of the non-linear structure and on the angles of incidence is given. Within the framework of the closed system under consideration it is shown that the imaginary part of the dielectric constant, determined by the value of the non-linear part of the polarisation at a frequency of the incident field, characterises the loss of energy in the non-linear medium which is spent for the generation of the third harmonic, where the contributions caused be the influence of the weak electromagnetic fields of diffraction are taken into account.

2. Maxwell equations and wave propagation in non-linear media with cubic polarisability

Electrodynamical and optical wave phenomena in charge- and current-free media can be described by the Maxwell equations

$$\begin{aligned} \nabla \times \mathbf{E} &= -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}, & \nabla \times \mathbf{H} &= \frac{1}{c}\frac{\partial \mathbf{D}}{\partial t}, \\ \nabla \cdot \mathbf{D} &= 0, & \nabla \cdot \mathbf{B} &= 0. \end{aligned} \tag{1}$$

Here $\mathbf{E} = \mathbf{E}(\mathbf{r},t)$, $\mathbf{H} = \mathbf{H}(\mathbf{r},t)$, $\mathbf{D} = \mathbf{D}(\mathbf{r},t)$ and $\mathbf{B} = \mathbf{B}(\mathbf{r},t)$ denote the vectors of electric and magnetic field intensities, electric displacement, and magnetic induction, respectively, and $(\mathbf{r},t) \in \mathbb{R}^3 \times (0,\infty)$. The symbol ∇ represents the formal vector of partial derivatives w.r.t. the spatial variables, i.e. $\nabla := (\partial/\partial x, \partial/\partial y, \partial/\partial z)^\top$, where the symbol $^\top$ denotes the transposition. In addition, the system (1) is completed by the material equations

$$\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P}, \qquad \mathbf{B} = \mathbf{H} + 4\pi\mathbf{M}, \tag{2}$$

where $\mathbf{P}$ and $\mathbf{M}$ are the vectors of the polarisation and magnetic moment, respectively. In general, the polarisation vector $\mathbf{P}$ is non-linear with respect to the intensity and non-local both in time and space.

In the present paper, the non-linear medium under consideration is located in the region Ω^{cl}, where Ω^{cl} denotes the closure of the set Ω defined by

$$\Omega := \{\mathbf{r} = (x,y,z)^\top \in \mathbb{R}^3 : \ |z| < 2\pi\delta\}$$

for some $\delta > 0$ being fixed. That is, the non-linear medium represents an infinite plate of thickness $4\pi\delta$.

As in the book Akhmediev & Ankevich (2003), the investigations will be restricted to non-linear media having a spatially non-local response function, i.e. the spatial dispersion is ignored (cf. Agranovich & Ginzburg (1966)). In this case, the polarisation vector can be expanded in terms of the electric field intensity as follows:

$$\mathbf{P} = \boldsymbol{\chi}^{(1)}\mathbf{E} + (\boldsymbol{\chi}^{(2)}\mathbf{E})\mathbf{E} + ((\boldsymbol{\chi}^{(3)}\mathbf{E})\mathbf{E})\mathbf{E} + \dots, \tag{3}$$

where $\boldsymbol{\chi}^{(1)}$, $\boldsymbol{\chi}^{(2)}$, $\boldsymbol{\chi}^{(3)}$ are the media susceptibility tensors of rank one, two and three, with components $\{\chi^{(1)}_{ij}\}_{i,j=1}^3$, $\{\chi^{(2)}_{ijk}\}_{i,j,k=1}^3$ and $\{\chi^{(3)}_{ijkl}\}_{i,j,k,l=1}^3$, respectively (see Butcher (1965)). In

the case of media which are invariant under the operations of inversion, reflection and rotation, in particular of isotropic media, the quadratic term disappears.
It is convenient to split $\mathbf{P}$ into its linear and non-linear parts as

$$\mathbf{P} = \mathbf{P}^{(L)} + \mathbf{P}^{(NL)},$$

where $\mathbf{P}^{(L)} := \boldsymbol{\chi}^{(1)}\mathbf{E}$. Similarly, with $\boldsymbol{\varepsilon} := \mathbf{I} + 4\pi\boldsymbol{\chi}^{(1)}$ and $\mathbf{D}^{(L)} := \boldsymbol{\varepsilon}\mathbf{E}$, where $\mathbf{I}$ denotes the identity in $\mathbb{C}^3$, the displacement field in (2) can be decomposed as

$$\mathbf{D} = \mathbf{D}^{(L)} + 4\pi\mathbf{P}^{(NL)}. \tag{4}$$

$\boldsymbol{\varepsilon}$ is the linear term of the permittivity tensor. Furthermore we assume that the media are non-magnetic, i.e

$$\mathbf{M} = 0, \tag{5}$$

so that

$$\mathbf{B} = \mathbf{H} \tag{6}$$

by (2). Resolving the equations (1), (4) and (6) with respect to $\mathbf{H}$, a single vector-valued equation results:

$$\nabla^2\mathbf{E} - \nabla(\nabla\cdot\mathbf{E}) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{D}^{(L)} - \frac{4\pi}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{P}^{(NL)} = 0. \tag{7}$$

Equation (7) is of rather general character and is used, together with the material equations (4), in electrodynamics and optics. In each particular case, specific assumptions are made that allow to simplify its form. For example, the second term in (7) may be ignored in a number of cases. One such case is the study of isotropic media considered here, where

$$\boldsymbol{\varepsilon} = \varepsilon^{(L)}\mathbf{I}$$

with a scalar, possibly complex-valued function $\varepsilon^{(L)}$. Then

$$\nabla\cdot(\boldsymbol{\varepsilon}\mathbf{E}) = \nabla\varepsilon^{(L)}\cdot\mathbf{E} + \varepsilon^{(L)}\nabla\cdot\mathbf{E}. \tag{8}$$

From (1), (4) and (8) we see that

$$0 = \nabla\cdot\mathbf{D} = \nabla\cdot(\boldsymbol{\varepsilon}\mathbf{E}) + 4\pi\nabla\cdot\mathbf{P}^{(NL)} = \nabla\varepsilon^{(L)}\cdot\mathbf{E} + \varepsilon^{(L)}\nabla\cdot\mathbf{E} + 4\pi\nabla\cdot\mathbf{P}^{(NL)},$$

hence

$$\nabla\cdot\mathbf{E} = -\frac{1}{\varepsilon^{(L)}}\nabla\varepsilon^{(L)}\cdot\mathbf{E} - \frac{4\pi}{\varepsilon^{(L)}}\nabla\cdot\mathbf{P}^{(NL)}. \tag{9}$$

In addition, if the media under consideration are transversely inhomogeneous w.r.t. z, i.e. $\varepsilon^{(L)} = \varepsilon^{(L)}(z)$, if the wave is linearly E-polarised, i.e. $\mathbf{E} = (E_1,0,0)^\top$, $\mathbf{H} = (0,H_2,H_3)^\top$, and if the electric field $\mathbf{E}$ is homogeneous w.r.t. the coordinate x, i.e.

$$\mathbf{E}(\mathbf{r},t) = (E_1(t;y,z),0,0)^\top, \tag{10}$$

then the first term of the r.h.s in (9) vanishes.
For linear media, in particular in $\mathbb{R}^3\setminus\Omega^{cl}$, the second term of the r.h.s in (9) is not present. In the layer medium the expansion (3) together with (10) implies that the vector $\mathbf{P}^{(NL)}$ has only one non-trivial component which is homogeneous w.r.t. x, i.e. $\mathbf{P}^{(NL)}(\mathbf{r},t) = (P_1(t;y,z),0,0)^\top$.

Then the second term of the r.h.s in (9) vanishes in Ω^{cl}, too. Consequently, the second term in equation (7) disappears completely, and we arrive at

$$\nabla^2\mathbf{E} - \frac{\varepsilon^{(L)}}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{E} - \frac{4\pi}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{P}^{(NL)} = 0, \tag{11}$$

where ∇^2 reduces to the Laplacian w.r.t. y and z, i.e. $\nabla^2 := \partial^2/\partial y^2 + \partial^2/\partial z^2$.
In this paper we consider resonance effects caused by the irradiation of a non-linear dielectric layer by an intense electromagnetic field of the frequency ω, where we also take into consideration the higher-order harmonics of the electromagnetic field. A mathematical model for the case of a weakly non-linear Kerr-type dielectric layer can be found in Yatsyk (2007); Shestopalov & Yatsyk (2007); Kravchenko & Yatsyk (2007).

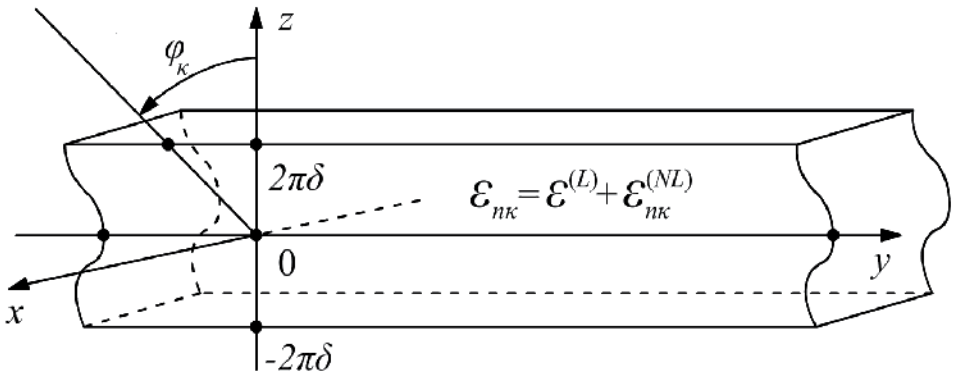


Fig. 1. The non-linear dielectric layered structure

The stationary problem of the diffraction of a plane electromagnetic wave (with oscillation frequency $\omega > 0$) on a transversely inhomogeneous, isotropic, non-magnetic, linearly polarised, dielectric layer filled with a cubically polarisable medium (see Fig. 1) is studied in the frequency domain (i.e. in the space of the Fourier amplitudes of the electromagnetic field), taking into account the multiple frequencies $s\omega$, $s \in \mathbb{N}$, of the excitation frequency generated by non-linear structure, where a time-dependence of the form $\exp(-is\omega t)$ is assumed. The transition between the time domain and the frequency domain is performed by means of direct and inverse Fourier transforms:

$$\hat{\mathbf{F}}(\mathbf{r},\hat{\omega}) = \int_{\mathbb{R}} \mathbf{F}(\mathbf{r},t)e^{i\hat{\omega}t}dt, \qquad \mathbf{F}(\mathbf{r},t) = \frac{1}{2\pi}\int_{\mathbb{R}} \hat{\mathbf{F}}(\mathbf{r},\hat{\omega})e^{-i\hat{\omega}t}d\hat{\omega},$$

where $\mathbf{F}$ is one of the vector fields $\mathbf{E}$ or $\mathbf{P}^{(NL)}$.
Applying formally the Fourier transform to equation (11), we obtain the following representation in the frequency domain:

$$\nabla^2\hat{\mathbf{E}}(\mathbf{r},\hat{\omega}) + \frac{\varepsilon^{(L)}\hat{\omega}^2}{c^2}\hat{\mathbf{E}}(\mathbf{r},\hat{\omega}) + \frac{4\pi\hat{\omega}^2}{c^2}\hat{\mathbf{P}}^{(NL)}(\mathbf{r},\hat{\omega}) = 0. \tag{12}$$

A stationary (i.e. $\sim \exp(-i\hat{\omega}t)$) electromagnetic wave propagating in a weakly non-linear dielectric structure gives rise to a field containing all frequency harmonics (see Agranovich & Ginzburg (1966), Vinogradova et al. (1990)). Therefore, the quantities describing the

electromagnetic field in the time domain subject to equation (11) can be represented as Fourier series

$$\mathbf{F}(\mathbf{r},t)=\frac{1}{2}\sum_{n\in\mathbb{Z}}\mathbf{F}(\mathbf{r},n\omega)e^{-in\omega t},\qquad \mathbf{F}\in\{\mathbf{E},\mathbf{P}^{(NL)}\}. \tag{13}$$

Applying to (13) the Fourier transform, we obtain

$$\hat{\mathbf{F}}(\mathbf{r},\hat{\omega})=\frac{1}{2}\int_{\mathbb{R}}\sum_{n\in\mathbb{Z}}\mathbf{F}(\mathbf{r},n\omega)e^{-in\omega t}e^{i\hat{\omega}t}dt=\frac{\sqrt{2\pi}}{2}\mathbf{F}(\mathbf{r},n\omega)\delta(\hat{\omega},n\omega),\qquad \mathbf{F}\in\{\mathbf{E},\mathbf{P}^{(NL)}\}, \tag{14}$$

where $\delta(s,s_0):=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}e^{i(s-s_0)t}dt$ is the Dirac delta-function located at $s=s_0$.
Substituting (14) into (12), we obtain an infinite system of equations with respect to the Fourier amplitudes of the electric field intensities of the non-linear structure in the frequency domain:

$$\nabla^2\mathbf{E}(\mathbf{r},s\omega)+\frac{\varepsilon^{(L)}(s\omega)^2}{c^2}\mathbf{E}(\mathbf{r},s\omega)+\frac{4\pi(s\omega)^2}{c^2}\mathbf{P}^{(NL)}(\mathbf{r},s\omega)=0,\quad s\in\mathbb{Z}. \tag{15}$$

For linear electrodynamic objects, the equations in system (15) are independent. In a non-linear structure, the presence of the functions $\mathbf{P}^{(NL)}(\mathbf{r},s\omega)$ makes them coupled since every harmonic depends on a series of $\mathbf{E}(\mathbf{r},s\omega)$. Indeed, consider a three-component E-polarised electromagnetic field $\mathbf{E}(\mathbf{r},s\omega)=(E_1(s\omega;y,z),0,0)^\top$, $\mathbf{H}(\mathbf{r},s\omega)=(0;H_2(s\omega;y,z),H_3(s\omega;y,z))^\top$. Since the field $\mathbf{E}$ has only one non-trivial component, the system (15) reduces to a system of scalar equations with respect to E_1:

$$\nabla^2E_1(\mathbf{r},s\omega)+\frac{\varepsilon^{(L)}(s\omega)^2}{c^2}E_1(\mathbf{r},s\omega)+\frac{4\pi(s\omega)^2}{c^2}P_1^{(NL)}(\mathbf{r},s\omega)=0,\quad s\in\mathbb{N}. \tag{16}$$

In writing equation (16), the property $E_1(\mathbf{r};j\omega)=E_1^*(\mathbf{r};-j\omega)$ of the Fourier coefficients and the lack of influence of the static electric field $E_1(\mathbf{r},s\omega)|_{s=0}=0$ on the non-linear structure were taken into consideration.
We assume that the main contribution to the non-linearity is introduced by the term $\mathbf{P}^{(NL)}(\mathbf{r},s\omega)$ (cf. Yatsyk (2007), Shestopalov & Yatsyk (2007), Kravchenko & Yatsyk (2007), Angermann & Yatsyk (2008), Yatsyk (2006), Schürmann et al. (2001), Smirnov et al. (2005), Serov et al. (2004)). We take only the lowest-order terms in the Taylor series expansion of the non-linear part $\mathbf{P}^{(NL)}(\mathbf{r},s\omega)=\left(P_1^{(NL)}(\mathbf{r},s\omega),0,0\right)^\top$ of the polarisation vector in the vicinity of the zero value of the electric field intensity, cf. (3). In this case, the only non-trivial component of the polarisation vector is determined by susceptibility tensor of the third order $\boldsymbol{\chi}^{(3)}$, that is characteristic for a non-linear isotropic medium with cubic polarisability. In the time domain, this component can be represented in the form (cf. (3) and (13)):

$$\begin{aligned}P_1^{(NL)}(\mathbf{r},t) &= \frac{1}{2}\sum_{s\in\mathbb{Z}\setminus\{0\}}P_1^{(NL)}(\mathbf{r},s\omega)e^{-is\omega t}\doteq\chi_{1111}^{(3)}E_1(\mathbf{r},t)E_1(\mathbf{r},t)E_1(\mathbf{r},t)\\ &\doteq \frac{1}{8}\sum_{\substack{n,m,p,s\in\mathbb{Z}\setminus\{0\}\\ n+m+p=s}}\chi_{1111}^{(3)}(s\omega;n\omega,m\omega,p\omega)E_1(\mathbf{r},n\omega)E_1(\mathbf{r},m\omega)E_1(\mathbf{r},p\omega)e^{-is\omega t},\end{aligned} \tag{17}$$

where the symbol $\dot{=}$ means that higher-order terms are neglected. Applying to (17) the Fourier transform with respect to time (14) we obtain an expansion in the frequency domain:

$$\begin{aligned} P_1^{(NL)}(\mathbf{r},s\omega) &= \frac{1}{4}\sum_{\substack{n,m,p\in\mathbb{Z}\setminus\{0\}\\ n+m+p=s}} \chi^{(3)}_{1111}(s\omega;n\omega,m\omega,p\omega)E_1(\mathbf{r},n\omega)E_1(\mathbf{r},m\omega)E_1(\mathbf{r},p\omega) \\ &= \frac{1}{4}\sum_{j\in\mathbb{N}} 3\chi^{(3)}_{1111}(s\omega;j\omega,-j\omega,s\omega)|E_1(\mathbf{r},j\omega)|^2E_1(\mathbf{r},s\omega) \\ &\quad +\frac{1}{4}\sum_{\substack{n,m,p\in\mathbb{Z}\setminus\{0\}\\ n\neq -m,\ p=s\\ m\neq -p,\ n=s\\ n\neq -p,\ m=s\\ n+m+p=s}} \chi^{(3)}_{1111}(s\omega;n\omega,m\omega,p\omega)E_1(\mathbf{r},n\omega)E_1(\mathbf{r},m\omega)E_1(\mathbf{r},p\omega). \end{aligned} \tag{18}$$

The addends in the first sum of the last representation of $P_1^{(NL)}(\mathbf{r},s\omega)$ in (18) are usually called *phase self-modulation (PSM) terms* (cf. Akhmediev & Ankevich (2003)). We denote them by $P_1^{(FSM)}(\mathbf{r},s\omega)$. Since the terms in formula (18) contain the factor $E_1(\mathbf{r},s\omega)$, they are responsible for the variation of the dielectric permittivity of the non-linear medium influenced by a variation of the amplitude of the field of excitation. We obtained them taking into account the property of the Fourier coefficients $E_1(\mathbf{r};j\omega) = E_1^*(\mathbf{r};-j\omega)$, where the factor 3 appears as a result of permutations $\{j\omega,-j\omega,s\omega\}$ of the three last parameters in the terms

$$\chi^{(3)}_{1111}(s\omega;j\omega,-j\omega,s\omega).$$

The addends in the second sum of the last representation of $P_1^{(NL)}(\mathbf{r},s\omega)$ in (18) are responsible for the generation of the multiple harmonics. Some of them generate radiation at multiple frequencies, others describe the mutual influence of the generated fields at multiple frequencies on the electromagnetic field being investigated. Moreover, those of them which clearly depend at the multiple frequency $s\omega$ on the unknown field of diffraction induce a complex contribution to the dielectric permittivity of the non-linear medium. They are denoted by $P_1^{(GC)}(\mathbf{r},s\omega)$. The remaining terms of the second sum are denoted by $P_1^{(G)}(\mathbf{r},s\omega)$. They play the role of the sources generating radiation. In summary, we have the representation

$$P_1^{(NL)}(\mathbf{r},s\omega) = P_1^{(FSM)}(\mathbf{r},s\omega) + P_1^{(GC)}(\mathbf{r},s\omega) + P_1^{(G)}(\mathbf{r},s\omega). \tag{19}$$

Thus, under the above assumption, the electromagnetic waves in a non-linear medium with a cubic polarisability are described by an infinite system (16)&(18) of non-linear equations (Yatsyk (2007), Shestopalov & Yatsyk (2007), Kravchenko & Yatsyk (2007), Angermann & Yatsyk (2010)). In what follows we will consider the equations in the frequency space taking into account the relation $\kappa = \frac{\omega}{c}$.

In the study of particular non-linear effects it proves to be possible to restrict the examination of the system (16)&(18) to a finite number of equations, and also to leave particular terms in the representation of the polarisation coefficients, which characterise the physical problem under investigation. For example, in the analysis of the non-linear effects caused by the generation of harmonics only at three combined frequencies (i.e., neglecting the influence of

higher harmonics), it is possible to restrict the investigation to a system of three equations. Taking into account only the non-trivial terms in the expansion of the polarisation coefficients, we arrive at the following system:

$$
\begin{cases}
\nabla^2 E_1(\mathbf{r},\kappa) + \varepsilon^{(L)}\kappa^2 E_1(\mathbf{r},\kappa) + 4\pi\kappa^2 P_1^{(NL)}(\mathbf{r},\kappa) &= 0,\\
\nabla^2 E_1(\mathbf{r},2\kappa) + \varepsilon^{(L)}(2\kappa)^2 E_1(\mathbf{r},2\kappa) + 4\pi(2\kappa)^2 P_1^{(NL)}(\mathbf{r},2\kappa) &= 0,\\
\nabla^2 E_1(\mathbf{r},3\kappa) + \varepsilon^{(L)}(3\kappa)^2 E_1(\mathbf{r},3\kappa) + 4\pi(3\kappa)^2 P_1^{(NL)}(\mathbf{r},3\kappa) &= 0,
\end{cases}
$$

$$
\begin{aligned}
P_1^{(NL)}(\mathbf{r},n\kappa) &= \frac{3}{4}\Big(\chi^{(3)}_{1111}(n\kappa;\kappa,-\kappa,n\kappa)|E_1(\mathbf{r},\kappa)|^2 + \chi^{(3)}_{1111}(n\kappa;2\kappa,-2\kappa,n\kappa)|E_1(\mathbf{r},2\kappa)|^2\\
&\quad + \chi^{(3)}_{1111}(n\kappa;3\kappa,-3\kappa,n\kappa)|E_1(\mathbf{r},3\kappa)|^2\Big)E_1(\mathbf{r},n\kappa)\\
&\quad + \delta_{n1}\frac{3}{4}\Big\{\chi^{(3)}_{1111}(\kappa;-\kappa,-\kappa,3\kappa)\left[E_1^*(\mathbf{r},\kappa)\right]^2 E_1(\mathbf{r},3\kappa)\\
&\qquad + \chi^{(3)}_{1111}(\kappa;2\kappa,2\kappa,-3\kappa)E_1^2(\mathbf{r},2\kappa)E_1^*(\mathbf{r},3\kappa)\Big\}\\
&\quad + \delta_{n2}\frac{3}{4}\chi^{(3)}_{1111}(2\kappa;-2\kappa,\kappa,3\kappa)E_1^*(\mathbf{r},2\kappa)E_1(\mathbf{r},\kappa)E_1(\mathbf{r},3\kappa)\\
&\quad + \delta_{n3}\left\{\frac{1}{4}\chi^{(3)}_{1111}(3\kappa;\kappa,\kappa,\kappa)E_1^3(\mathbf{r},\kappa) + \frac{3}{4}\chi^{(3)}_{1111}(3\kappa;2\kappa,2\kappa,-\kappa)E_1^2(\mathbf{r},2\kappa)E_1^*(\mathbf{r},\kappa)\right\},\\
&\qquad\qquad n=1,2,3,
\end{aligned}
\tag{20}
$$

where δ_{nm} denotes Kronecker's symbol. Using (19), we obtain

$$
\begin{cases}
\nabla^2 E_1(\mathbf{r},\kappa) + \varepsilon^{(L)}\kappa^2 E_1(\mathbf{r},\kappa) + 4\pi\kappa^2\left(P_1^{(FSM)}(\mathbf{r},\kappa) + P_1^{(GC)}(\mathbf{r},\kappa)\right)\\
\qquad\qquad = -4\pi\kappa^2 P_1^{(G)}(\mathbf{r},\kappa),\\
\nabla^2 E_1(\mathbf{r},2\kappa) + \varepsilon^{(L)}(2\kappa)^2 E_1(\mathbf{r},2\kappa) + 4\pi(2\kappa)^2\left(P_1^{(FSM)}(\mathbf{r},2\kappa) + P_1^{(GC)}(\mathbf{r},2\kappa)\right) = 0,\\
\nabla^2 E_1(\mathbf{r},3\kappa) + \varepsilon^{(L)}(3\kappa)^2 E_1 r(\mathbf{r},3\kappa) + 4\pi(3\kappa)^2 P_1^{(FSM)}(\mathbf{r},3\kappa)\\
\qquad\qquad = -4\pi(3\kappa)^2 P_1^{(G)}(\mathbf{r},3\kappa),
\end{cases}
$$

$$
\begin{aligned}
P_1^{(FSM)}(\mathbf{r},n\kappa) &= \frac{3}{4}(\chi^{(3)}_{1111}(n\kappa;\kappa,-\kappa,n\kappa)|E_1(\mathbf{r},\kappa)|^2 + \chi^{(3)}_{1111}(n\kappa;2\kappa,-2\kappa,n\kappa)|E_1(\mathbf{r},2\kappa)|^2\\
&\quad + \chi^{(3)}_{1111}(n\kappa;3\kappa,-3\kappa,n\kappa)|E_1(\mathbf{r},3\kappa)|^2)E_1(\mathbf{r},n\kappa),\quad n=1,2,3,\\
P_1^{(GC)}(\mathbf{r},\kappa) &= \frac{3}{4}\chi^{(3)}_{1111}(\kappa;-\kappa,-\kappa,3\kappa)\left[E_1^*(\mathbf{r},\kappa)\right]^2 E_1(\mathbf{r},3\kappa)\\
&= \frac{3}{4}\chi^{(3)}_{1111}(\kappa;-\kappa,-\kappa,3\kappa)\frac{\left[E_1^*(\mathbf{r},\kappa)\right]^2}{E_1(\mathbf{r},\kappa)}E_1(\mathbf{r},3\kappa)E_1(\mathbf{r},\kappa),\\
P_1^{(G)}(\mathbf{r},\kappa) &= \frac{3}{4}\chi^{(3)}_{1111}(\kappa;2\kappa,2\kappa,-3\kappa)E_1^2(\mathbf{r},2\kappa)E_1^*(\mathbf{r},3\kappa),\\
P_1^{(GC)}(\mathbf{r},2\kappa) &= \frac{3}{4}\chi^{(3)}_{1111}(2\kappa;-2\kappa,\kappa,3\kappa)E_1^*(\mathbf{r},2\kappa)E_1(\mathbf{r},\kappa)E_1(\mathbf{r},3\kappa)\\
&= \frac{3}{4}\chi^{(3)}_{1111}(2\kappa;-2\kappa,\kappa,3\kappa)\frac{E_1^*(\mathbf{r},2\kappa)}{E_1(\mathbf{r},2\kappa)}E_1(\mathbf{r},\kappa)E_1(\mathbf{r},3\kappa)E_1(\mathbf{r},2\kappa),\\
P_1^{(G)}(\mathbf{r},3\kappa) &= \frac{3}{4}\left\{\frac{1}{3}\chi^{(3)}_{1111}(3\kappa;\kappa,\kappa,\kappa)E_1^3(\mathbf{r},\kappa) + \chi^{(3)}_{1111}(3\kappa;2\kappa,2\kappa,-\kappa)E_1^2(\mathbf{r},2\kappa)E_1^*(\mathbf{r},\kappa)\right\}.
\end{aligned}
\tag{21}
$$

The analysis of the problem can be significantly simplified by reducing the number of parameters, i.e. the coefficients of the cubic susceptibility of the non-linear medium. Thus,

by Kleinman's rule (Kleinman (1962), Miloslavski (2008)),

$$\begin{array}{ccccccc} & \chi^{(3)}_{1111}(n\kappa;\kappa,-\kappa,n\kappa) & = & \chi^{(3)}_{1111}(n\kappa;2\kappa,-2\kappa,n\kappa) & = & \chi^{(3)}_{1111}(n\kappa;3\kappa,-3\kappa,n\kappa) \\ = & \chi^{(3)}_{1111}(\kappa;-\kappa,-\kappa,3\kappa) & = & \chi^{(3)}_{1111}(\kappa;2\kappa,2\kappa,-3\kappa) & = & \chi^{(3)}_{1111}(2\kappa;-2\kappa,\kappa,3\kappa) \\ = & \chi^{(3)}_{1111}(3\kappa;\kappa,\kappa,\kappa) & = & \chi^{(3)}_{1111}(3\kappa;2\kappa,2\kappa,-\kappa) & =: & \chi^{(3)}_{1111}, \quad n=1,2,3. \end{array}$$

Therefore, the system (21) can be written in the form

$$\begin{cases} \nabla^2 E_1(\mathbf{r},\kappa)+\varepsilon^{(L)}\kappa^2 E_1(\mathbf{r},\kappa)+4\pi\kappa^2\left(P_1^{(FSM)}(\mathbf{r},\kappa)+P_1^{(GC)}(\mathbf{r},\kappa)\right) \\ \qquad\qquad = -4\pi\kappa^2 P_1^{(G)}(\mathbf{r},\kappa), \\ \nabla^2 E_1(\mathbf{r},2\kappa)+\varepsilon^{(L)}(2\kappa)^2 E_1(\mathbf{r},2\kappa)+4\pi(2\kappa)^2\left(P_1^{(FSM)}(\mathbf{r},2\kappa)+P_1^{(GC)}(\mathbf{r},2\kappa)\right) = 0, \\ \nabla^2 E_1(\mathbf{r},3\kappa)+\varepsilon^{(L)}(3\kappa)^2 E_1(\mathbf{r},3\kappa)+4\pi(3\kappa)^2 P_1^{(FSM)}(\mathbf{r},3\kappa) \\ \qquad\qquad = -4\pi(3\kappa)^2 P_1^{(G)}(\mathbf{r},3\kappa), \end{cases}$$

$$\begin{array}{rcl} P_1^{(FSM)}(\mathbf{r},n\kappa) & = & \frac{3}{4}\chi^{(3)}_{1111}(|E_1(\mathbf{r},\kappa)|^2+|E_1(\mathbf{r},2\kappa)|^2+|E_1(\mathbf{r},3\kappa)|^2)E_1(\mathbf{r},n\kappa), \quad n=1,2,3, \\ P_1^{(GC)}(\mathbf{r},\kappa) & = & \frac{3}{4}\chi^{(3)}_{1111}\frac{\left[E_1^*(\mathbf{r},\kappa)\right]^2}{E_1(\mathbf{r},\kappa)}E_1(\mathbf{r},3\kappa)E_1(\mathbf{r},\kappa), \\ P_1^{(G)}(\mathbf{r},\kappa) & = & \frac{3}{4}\chi^{(3)}_{1111}E_1^2(\mathbf{r},2\kappa)E_1^*(\mathbf{r},3\kappa), \\ P_1^{(GC)}(\mathbf{r},2\kappa) & = & \frac{3}{4}\chi^{(3)}_{1111}\frac{E_1^*(\mathbf{r},2\kappa)}{E_1(\mathbf{r},2\kappa)}E_1(\mathbf{r},\kappa)E_1(\mathbf{r},3\kappa)E_1(\mathbf{r},2\kappa), \quad P_1^{(G)}(\mathbf{r},2\kappa):=0, \\ P_1^{(G)}(\mathbf{r},3\kappa), & = & \frac{3}{4}\chi^{(3)}_{1111}\left\{\frac{1}{3}E_1^3(\mathbf{r},\kappa)+E_1^2(\mathbf{r},2\kappa)E_1^*(\mathbf{r},\kappa)\right\}, \quad P_1^{(GC)}(\mathbf{r},3\kappa):=0. \end{array} \tag{22}$$

The permittivity of the non-linear medium filling a layer (see Fig. 1) can be represented as

$$\varepsilon_{n\kappa}=\varepsilon^{(L)}+\varepsilon^{(NL)}_{n\kappa} \quad \text{for} \quad |z|\le 2\pi\delta. \tag{23}$$

Outside the layer, i.e. for $|z|>2\pi\delta$, $\varepsilon_{n\kappa}=1$. The linear and non-linear terms of the permittivity of the layer are given by the coefficients at $(n\kappa)^2E_1(\mathbf{r},n\kappa)$ in the second and third addends in each of the equations of the system, respectively. Thus

$$\varepsilon^{(L)}=\frac{D_1^{(L)}(\mathbf{r},n\kappa)}{E_1(\mathbf{r},n\kappa)}=1+4\pi\chi^{(1)}_{11}, \tag{24}$$

where the representations for the linear part of the complex components of the electric displacement $D_1^{(L)}(\mathbf{r},n\kappa)=E_1(\mathbf{r},n\kappa)+4\pi P_1^{(L)}(\mathbf{r},n\kappa)=\varepsilon^{(L)}E_1(\mathbf{r},n\kappa)$ and the polarisation $P_1^{(L)}(\mathbf{r},n\kappa)=\chi^{(1)}_{11}E_1(\mathbf{r},n\kappa)$ are taken into account. Similarly, the third term of each equation of the system makes it possible to write the non-linear component of the permittivity in the form

$$\begin{array}{rcl} \varepsilon^{(NL)}_{n\kappa} & = & 4\pi\dfrac{P_1^{(FSM)}(\mathbf{r},n\kappa)+P_1^{(GC)}(\mathbf{r},n\kappa)}{E_1(\mathbf{r},n\kappa)} \\ & = & \alpha(z)\big[|E_1(\mathbf{r},\kappa)|^2+|E_1(\mathbf{r},2\kappa)|^2+|E_1(\mathbf{r},3\kappa)|^2 \\ & & +\delta_{n1}\dfrac{\left[E_1^*(\mathbf{r},\kappa)\right]^2}{E_1(\mathbf{r},\kappa)}E_1(\mathbf{r},3\kappa)+\delta_{n2}\dfrac{E_1^*(\mathbf{r},2\kappa)}{E_1(\mathbf{r},2\kappa)}E_1(\mathbf{r},\kappa)E_1(\mathbf{r},3\kappa)\big], \end{array} \tag{25}$$

where $\alpha(z) := 3\pi\chi^{(3)}_{1111}(z)$ is the so-called function of the cubic susceptibility of the non-linear medium.
For transversely inhomogeneous media (a layer or a layered structure), the linear part $\varepsilon^{(L)} = \varepsilon^{(L)}(z) = 1 + 4\pi\chi^{(1)}_{11}(z)$ of the permittivity (cf. (24)) is described by a piecewise smooth or a piecewise constant function. Similarly, the function of the cubic susceptibility $\alpha = \alpha(z)$ is also a piecewise smooth or a piecewise constant function. This assumption allows us to investigate the diffraction characteristics of a non-linear layer and of a layered structure (consisting of a finite number of non-linear dielectric layers) within one and the same mathematical model.

3. The condition of phase synchronism. Quasi-homogeneous electromagnetic fields in a transversely inhomogeneous non-linear dielectric layered structure.

The scattered and generated field in a transversely inhomogeneous, non-linear dielectric layer excited by a plane wave is quasi-homogeneous along the coordinate y, hence it can be represented as

(C1) $E_1(\mathbf{r}, n\kappa) =: E_1(n\kappa; y, z) := U(n\kappa; z)\exp(i\phi_{n\kappa} y), \quad n = 1,2,3.$

Here $U(n\kappa; z)$ and $\phi_{n\kappa} := n\kappa\sin\varphi_{n\kappa}$ denote the complex-valued transverse component of the Fourier amplitude of the electric field and the value of the longitudinal propagation constant (longitudinal wave-number) at the frequency $n\kappa$, respectively, where $\varphi_{n\kappa}$ is the given angle of incidence of the exciting field of frequency $n\kappa$ (cf. Fig. 1).
The dielectric permittivities of the layered structure at the multiple frequencies $n\kappa$ are determined by the values of the transverse components of the Fourier amplitudes of the scattered and generated fields, i.e. by the redistribution of energy of the electric fields at multiple frequencies, where the angles of incidence are given and the non-linear structure under consideration is transversely inhomogeneous. The condition of the longitudinal homogeneity (along the coordinate y) of the non-linear dielectric constant of the layered structure can be written as

$$\varepsilon^{(NL)}_{n\kappa}(z, \alpha(z), E_1(\mathbf{r},\kappa), E_1(\mathbf{r},2\kappa), E_1(\mathbf{r},3\kappa)) = \varepsilon^{(NL)}_{n\kappa}(z, \alpha(z), U(\kappa;z), U(2\kappa;z), U(3\kappa;z)). \tag{26}$$

Using the representation (25) and the conditions (C1), (26), we obtain the following physically consistent requirement, which we call *the condition of the phase synchronism of waves:*

(C2) $\phi_{n\kappa} = n\phi_\kappa, \quad n = 1,2,3.$

Indeed, from (25) and (C1) it follows that

$$\begin{aligned}
\varepsilon^{(NL)}_{n\kappa} &= \alpha(z)\big[|E_1(\mathbf{r},\kappa)|^2 + |E_1(\mathbf{r},2\kappa)|^2 + |E_1(\mathbf{r},3\kappa)|^2 \\
&\qquad + \delta_{n1}\frac{\left[E_1^*(\mathbf{r},\kappa)\right]^2}{E_1(\mathbf{r},\kappa)}E_1(\mathbf{r},3\kappa) + \delta_{n2}\frac{E_1^*(\mathbf{r},2\kappa)}{E_1(\mathbf{r},2\kappa)}E_1(\mathbf{r},\kappa)E_1(\mathbf{r},3\kappa)\big], \\
&= \alpha(z)\big[|U(\kappa;z)|^2 + |U(2\kappa;z)|^2 + |U(3\kappa;z)|^2 \\
&\qquad + \delta_{n1}\frac{\left[U^*(\mathbf{r},\kappa)\right]^2}{U(\mathbf{r},\kappa)}U(3\kappa;z)\exp\{i[-3\phi_\kappa + \phi_{3\kappa}]y\} \\
&\qquad + \delta_{n2}\frac{U^*(\mathbf{r},2\kappa)}{U(\mathbf{r},2\kappa)}U(\kappa;z)U(3\kappa;z)\exp\{i[-2\phi_{2\kappa} + \phi_\kappa + \phi_{3\kappa}]y\}\big], \quad n = 1,2,3.
\end{aligned} \tag{27}$$

Therefore the condition (26) is satisfied if

$$\begin{cases} -3\phi_\kappa + \phi_{3\kappa} = 0, \\ -2\phi_{2\kappa} + \phi_\kappa + \phi_{3\kappa} = 0. \end{cases} \tag{28}$$

From this system we obtain the condition (C2).
According to (23), (24), (27) and (C2), the permittivity of the non-linear layer can be expressed as

$$\begin{aligned}
&\varepsilon_{n\kappa}(z,\alpha(z),E_1(\mathbf{r},\kappa),E_1(\mathbf{r},2\kappa),E_1(\mathbf{r},3\kappa))\\
=\;&\varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\\
=\;&\varepsilon^{(L)}(z)+\alpha(z)\big[|U(\kappa;z)|^2+|U(2\kappa;z)|^2+|U(3\kappa;z)|^2\\
&+\delta_{n1}U^*(\kappa;z)\exp\{-2i\arg(U(\kappa;z))\}U(3\kappa;z)\\
&+\delta_{n2}\exp\{-2i\arg(U(2\kappa;z))\}U(\kappa,z)U(3\kappa;z)\big]\\
=\;&\varepsilon^{(L)}(z)+\alpha(z)\big[|U(\kappa;z)|^2+|U(2\kappa;z)|^2+|U(3\kappa;z)|^2\\
&+\delta_{n1}|U(\kappa;z)||U(3\kappa;z)|\exp\{i[-3\arg(U(\kappa;z))-3\phi_\kappa y+\arg(U(3\kappa;z))+\phi_{3\kappa}y]\}\\
&+\delta_{n2}|U(\kappa;z)||U(3\kappa;z)|\exp\{i[-2\arg(U(2\kappa;z))-2\phi_{2\kappa}y+\arg(U(\kappa;z))+\phi_\kappa y\\
&\qquad+\arg(U(3\kappa;z))+\phi_{3\kappa}y]\}\big]\\
=\;&\varepsilon^{(L)}(z)+\alpha(z)\big[|U(\kappa;z)|^2+|U(2\kappa;z)|^2+|U(3\kappa;z)|^2\\
&+\delta_{n1}|U(\kappa;z)||U(3\kappa;z)|\exp\{i[-3\arg(U(\kappa;z))+\arg(U(3\kappa;z))]\}\\
&+\delta_{n2}|U(\kappa;z)||U(3\kappa;z)|\exp\{i[-2\arg(U(2\kappa;z))+\arg(U(\kappa;z))+\arg(U(3\kappa;z))]\}\big],\\
&\qquad\qquad n=1,2,3.
\end{aligned}\tag{29}$$

The investigation of the quasi-homogeneous fields $E_1(n\kappa;y,z)$ (cf. condition (C1)) in a transversely inhomogeneous non-linear dielectric layer shows that, if the condition of the phase synchronism (C2) is satisfied, the components of the non-linear polarisation $P_1^{(G)}(\mathbf{r},n\kappa)$ (playing the role of the sources generating radiation in the right-hand sides of the system (22)) satisfy the quasi-homogeneity condition, too. Indeed, using (25) and (C1), the right-hand sides of the first and third equations of (22) can be rewritten as

$$\begin{aligned}
-4\pi\kappa^2P_1^{(G)}(\mathbf{r},\kappa) &= -\alpha(z)\kappa^2E_1^2(\mathbf{r},2\kappa)E_1^*(\mathbf{r},3\kappa)\\
&= -\alpha(z)\kappa^2U^2(2\kappa;z)U^*(3\kappa;z)\exp\{i[2\phi_{2\kappa}-\phi_{3\kappa}]y\}\\
&= -\alpha(z)\kappa^2U^2(2\kappa;z)U^*(3\kappa;z)\exp(i\phi_\kappa y)
\end{aligned}$$

and

$$\begin{aligned}
-4\pi(3\kappa)^2P_1^{(G)}(\mathbf{r},3\kappa) &= -\alpha(z)(3\kappa)^2\left\{\frac{1}{3}E_1^3(\mathbf{r},\kappa)+E_1^2(\mathbf{r},2\kappa)E_1^*(\mathbf{r},\kappa)\right\}\\
&= -\alpha(z)(3\kappa)^2\left\{\frac{1}{3}U^3(\kappa;z)\exp(3i\phi_\kappa y)\right.\\
&\qquad\left.+U^2(2\kappa;z)U^*(\kappa;z)\exp\{i[2\phi_{2\kappa}-\phi_\kappa]y\}\right\}\\
&= -\alpha(z)(3\kappa)^2\left\{\frac{1}{3}U^3(\kappa;z)+U^2(2\kappa;z)U^*(\kappa;z)\right\}\exp(i\phi_{3\kappa}y),
\end{aligned}$$

respectively. This shows that the quasi-homogeneity condition for the components of the non-linear polarisation $P_1^{(G)}(\mathbf{r},n\kappa)$ is satisfied.
In the considered case of spatially quasi-homogeneous (along the coordinate y) electromagnetic fields (C1), the condition of the phase synchronism of waves (C2) reads as

$$\sin\varphi_{n\kappa}=\sin\varphi_\kappa,\quad n=1,2,3.$$

Consequently, the given angle of incidence of a plane wave at the frequency κ coincides with the possible directions of the angles of incidence of plane waves at the multiple frequencies $n\kappa$. The angles of the wave scattered by the layer are equal to $\varphi_{n\kappa}^{\text{scat}}=-\varphi_{n\kappa}$ in the zone of

reflection $z > 2\pi\delta$ and $\varphi_{n\kappa}^{\text{scat}} = \pi + \varphi_{n\kappa}$ and in the zone of transmission of the non-linear layer $z < -2\pi\delta$, where all angles are measured counter-clockwise in the (y,z)-plane from the z-axis (cf. Fig. 1).

4. The diffraction of a packet of plane waves on a non-linear layered dielectric structure. The third harmonics generation

As a first observation we mention that the effect of a weak quasi-homogeneous electromagnetic field (C1) on the non-linear dielectric structure such that harmonics at multiple frequencies are not generated, i.e. $E_1(\mathbf{r},2\kappa) = 0$ and $E_1(\mathbf{r},3\kappa) = 0$, reduces to find the electric field component $E_1(\mathbf{r},\kappa)$ determined by the first equation of the system (22). In this case, a diffraction problem for a plane wave on a non-linear dielectric layer with a Kerr-type non-linearity $\varepsilon_{n\kappa} = \varepsilon^{(L)}(z) + \alpha(z)|E_1(\mathbf{r},\kappa)|^2$ and a vanishing right-hand side is to be solved, see Yatsyk (2007); Shestopalov & Yatsyk (2007); Kravchenko & Yatsyk (2007); Angermann & Yatsyk (2008); Yatsyk (2006); Smirnov et al. (2005); Serov et al. (2004).

The generation process of a field at the triple frequency 3κ by the non-linear dielectric structure is caused by a strong incident electromagnetic field at the frequency κ and can be described by the first and third equations of the system (22) only. Since the right-hand side of the second equation in (22) is equal to zero, we may set $E_1(\mathbf{r},2\kappa) = 0$ corresponding to the homogeneous boundary condition w.r.t. $E_1(\mathbf{r},2\kappa)$. Therefore the second equation in (22) can be completely omitted.

A further interesting problem consists in the investigation of the influence of a packet of waves on the generation of the third harmonic, if a strong incident field at the basic frequency κ and, in addition, weak incident quasi-homogeneous electromagnetic fields at the double and triple frequencies 2κ, 3κ (which alone do not generate harmonics at multiple frequencies) excite the non-linear structure. The system (22) allows to describe the corresponding process of the third harmonics generation. Namely, if such a wave packet consists of a strong field at the basic frequency κ and of a weak field at the triple frequency 3κ, then we arrive, as in the situation described above, at the system (22) with $E_1(\mathbf{r},2\kappa) = 0$, i.e. it is sufficient to consider the first and third equations of (22) only. For wave packets consisting of a strong field at the basic frequency κ and of a weak field at the frequency 2κ, (or of two weak fields at the frequencies 2κ and 3κ) we have to take into account all three equations of system (22). This is caused by the inhomogeneity of the corresponding diffraction problem, where a weak incident field at the double frequency 2κ (or two weak fields at the frequencies 2κ and 3κ) excites (resp. excite) the dielectric medium.

So we consider the problem of diffraction of a packet of plane waves consisting of a strong field at the frequency κ (which generates a field at the triple frequency 3κ) and of weak fields at the frequencies 2κ and 3κ (having an impact on the process of third harmonic generation due to the contribution of weak electromagnetic fields of diffraction)

$$\left\{E_1^{\text{inc}}(\mathbf{r},\kappa) := E_1^{\text{inc}}(\kappa;y,z) := a_{n\kappa}^{\text{inc}} \exp\left(i\big(\phi_{n\kappa}y - \Gamma_{n\kappa}(z - 2\pi\delta)\big)\right)\right\}_{n=1}^{3}, \quad z > 2\pi\delta, \tag{30}$$

with amplitudes $a_{n\kappa}^{\text{inc}}$ and angles of incidence $\varphi_{n\kappa}$, $|\varphi| < \pi/2$ (cf. Fig. 1), where $\phi_{n\kappa} := n\kappa \sin\varphi_{n\kappa}$ are the longitudinal propagation constants (longitudinal wave-numbers) and $\Gamma_{n\kappa} := \sqrt{(n\kappa)^2 - \phi_{n\kappa}^2}$ are the transverse propagation constants (transverse wave-numbers).

In this setting, the complex amplitudes of the total fields of diffraction

$$E_1(\mathbf{r},n\kappa) =: E_1(n\kappa;y,z) := U(n\kappa;z)\exp(i\phi_{n\kappa}y) := E_1^{\text{inc}}(n\kappa;y,z) + E_1^{\text{scat}}(n\kappa;y,z)$$

of a plane wave (30) in a non-magnetic, isotropic, linearly polarised

$$\begin{aligned}\mathbf{E}(\mathbf{r},n\kappa) &= (E_1(n\kappa;y,z),0,0)^\top,\\ \mathbf{H}(\mathbf{r},n\kappa) &= \left(0,\frac{1}{in\omega\mu_0}\frac{\partial E_1(n\kappa;y,z)}{\partial z},-\frac{1}{in\omega\mu_0}\frac{\partial E_1(n\kappa;y,z)}{\partial y}\right)^\top\end{aligned}$$

(E-polarisation), transversely inhomogeneous $\varepsilon^{(L)}=\varepsilon^{(L)}(z)=1+4\pi\chi_{11}^{(1)}(z)$ dielectric layer (see Fig. 1) with a cubic polarisability $\mathbf{P}^{(NL)}(\mathbf{r},n\kappa)=(P_1^{(NL)}(n\kappa;y,z),0,0)^\top$ of the medium (see (20)) satisfies the system of equations (cf. (22) – (25))

$$\left\{\begin{array}{rcl}\nabla^2E_1(\mathbf{r},\kappa)+\kappa^2\varepsilon_\kappa(z,\alpha(z),E_1(\mathbf{r},\kappa),E_1(\mathbf{r},2\kappa),E_1(\mathbf{r},3\kappa))E_1(\mathbf{r},\kappa) & & \\ \qquad = -\alpha(z)\kappa^2E_1^2(\mathbf{r},2\kappa)E_1^*(\mathbf{r},3\kappa), & & \\ \nabla^2E_1(\mathbf{r},2\kappa)+(2\kappa)^2\varepsilon_{2\kappa}(z,\alpha(z),E_1(\mathbf{r},\kappa),E_1(\mathbf{r},2\kappa),E_1(\mathbf{r},3\kappa))E_1(\mathbf{r},2\kappa) & = & 0,\\ \nabla^2E_1(\mathbf{r},3\kappa)+(3\kappa)^2\varepsilon_{3\kappa}(z,\alpha(z),E_1(\mathbf{r},\kappa),E_1(\mathbf{r},2\kappa),E_1(\mathbf{r},3\kappa))E_1(\mathbf{r},3\kappa) & & \\ \qquad = -\alpha(z)(3\kappa)^2\left\{\frac{1}{3}E_1^3(\mathbf{r},\kappa)+E_1^2(\mathbf{r},2\kappa)E_1^*(\mathbf{r},\kappa)\right\} & & \end{array}\right. \tag{31}$$

together with the following conditions, where $\mathbf{E}_{\rm tg}(n\kappa;y,z)$ and $\mathbf{H}_{\rm tg}(n\kappa;y,z)$ denote the tangential components of the intensity vectors of the full electromagnetic field $\{\mathbf{E}(n\kappa;y,z)\}_{n=1,2,3}$, $\{\mathbf{H}(n\kappa;y,z)\}_{n=1,2,3}$:

- (C1) $E_1(n\kappa;y,z)=U(n\kappa;z)\exp(i\phi_{n\kappa}y),\quad n=1,2,3$
  (the quasi-homogeneity condition w.r.t. the spatial variable y introduced in Section 3),
- (C2) $\phi_{n\kappa}=n\phi_\kappa,\quad n=1,2,3,$
  (the condition of phase synchronism of waves introduced in Section 3),
- (C3) $\mathbf{E}_{\rm tg}(n\kappa;y,z)$ and $\mathbf{H}_{\rm tg}(n\kappa;y,z)$ (i.e. $E_1(n\kappa;y,z)$ and $H_2(n\kappa;y,z)$) are continuous at the boundary layers of the non-linear structure,
- (C4) $E_1^{\rm scat}(n\kappa;y,z)=\left\{\begin{matrix}a_{n\kappa}^{\rm scat}\\ b_{n\kappa}^{\rm scat}\end{matrix}\right\}\exp\left(i\left(\phi_{n\kappa}y\pm\Gamma_{n\kappa}(z\mp2\pi\delta)\right)\right),\quad z\gtrless\pm2\pi\delta,\quad n=1,2,3$
  (the radiation condition w.r.t. the scattered field).

The condition (C4) provides a physically consistent behaviour of the energy characteristics of scattering and guarantees the absence of waves coming from infinity (i.e. $z=\pm\infty$), see Shestopalov & Sirenko (1989). We study the scattering properties of the non-linear layer, where in (C4) we always have $\Im\mathfrak{m}\,\Gamma_{n\kappa}=0$, $\Re\mathfrak{e}\,\Gamma_{n\kappa}>0$. Note that (C4) is also applicable for the analysis of the wave-guide properties of the layer, where $\Im\mathfrak{m}\,\Gamma_{n\kappa}>0$, $\Re\mathfrak{e}\,\Gamma_{n\kappa}=0$.
Here and in what follows we use the following notation: $(\mathbf{r},t)$ are dimensionless spatial-temporal coordinates such that the thickness of the layer is equal to $4\pi\delta$. The time-dependence is determined by the factors $\exp(-in\omega t)$, where $\omega:=\kappa c$ is the dimensionless circular frequency and κ is a dimensionless frequency parameter such that $\kappa=\omega/c:=2\pi/\lambda$. This parameter characterises the ratio of the true thickness h of the layer to the free-space wavelength λ, i.e. $h/\lambda=2\kappa\delta$. $c=(\varepsilon_0\mu_0)^{-1/2}$ denotes a dimensionless parameter, equal to the absolute value of the speed of light in the medium containing the layer ($\Im\mathfrak{m}\,c=0$). ε_0 and μ_0 are the material parameters of the medium. The absolute values of the true variables $\mathbf{r}',t',\omega'$ are given by the formulas $\mathbf{r}'=h\mathbf{r}/4\pi\delta$, $t'=th/4\pi\delta$, $\omega'=\omega4\pi\delta/h$.

The desired solution of the diffraction problem (31), (C1) – (C4) can be represented as follows:

$$\begin{array}{l} E_1(n\kappa;y,z) = U(n\kappa;z)\exp(i\phi_{n\kappa}y) \\ = \begin{cases} a^{\rm inc}_{n\kappa}\exp(i(\phi_{n\kappa}y - \Gamma_{n\kappa}(z-2\pi\delta))) + a^{\rm scat}_{n\kappa}\exp(i(\phi_{n\kappa}y + \Gamma_{n\kappa}(z-2\pi\delta))), & z > 2\pi\delta, \\ U(n\kappa;z)\exp(i\phi_{n\kappa}y), & |z| \le 2\pi\delta, \\ b^{\rm scat}_{n\kappa}\exp(i(\phi_{n\kappa}y - \Gamma_{n\kappa}(z+2\pi\delta))), & z < -2\pi\delta, \end{cases} \\ n = 1,2,3. \end{array} \tag{32}$$

Note that depending on the magnitudes of the amplitudes $\{a^{\rm inc}_{\kappa}, a^{\rm inc}_{2\kappa}, a^{\rm inc}_{3\kappa}\}$ of the packet of incident plane waves, the amplitudes $\{a^{\rm scat}_{n\kappa}, b^{\rm scat}_{n\kappa}\}_{n=1}^{3}$ of the scattered fields can be considered as the amplitudes of the diffraction field, of the generation field or of the sum of the diffraction and generation fields. If the components $\{a^{\rm inc}_{\kappa} = a^{\rm inc(w)}_{\kappa}, a^{\rm inc}_{2\kappa} = a^{\rm inc(w)}_{2\kappa}, a^{\rm inc}_{3\kappa} = a^{\rm inc(w)}_{3\kappa}\}$ of the packet consist of the amplitudes of weak fields, then $\{a^{\rm scat}_{n\kappa} = a^{\rm dif}_{n\kappa}, b^{\rm scat}_{n\kappa} = b^{\rm dif}_{n\kappa}\}_{n=1}^{3}$.
The presence of an amplitude of a strong field at the basic frequency κ in the packet $\{a^{\rm inc}_{\kappa} = a^{\rm inc(s)}_{\kappa}, a^{\rm inc}_{2\kappa} = a^{\rm inc(w)}_{2\kappa}, a^{\rm inc}_{3\kappa} = a^{\rm inc(w)}_{3\kappa}\}$ leads to non-trivial right-hand sides in the problem (31), (C1) – (C4). In this case the analysis of the following situations is of interest (see (32)):

$$\left.\begin{array}{l} a^{\rm inc}_{\kappa} = a^{\rm inc(s)}_{\kappa} \neq 0, \\ a^{\rm inc}_{2\kappa} = a^{\rm inc(w)}_{2\kappa} := 0, \\ a^{\rm inc}_{3\kappa} = a^{\rm inc(w)}_{3\kappa} := 0 \end{array}\right\} \quad \Rightarrow \quad \left\{\begin{array}{lll} a^{\rm scat}_{\kappa} = a^{\rm dif}_{\kappa}, & a^{\rm scat}_{2\kappa} = 0, & a^{\rm scat}_{3\kappa} = a^{\rm gen}_{3\kappa} \\ b^{\rm scat}_{\kappa} = b^{\rm dif}_{\kappa}, & b^{\rm scat}_{2\kappa} = 0, & b^{\rm scat}_{3\kappa} = b^{\rm gen}_{3\kappa} \end{array}\right\},$$

$$\left.\begin{array}{l} a^{\rm inc}_{\kappa} = a^{\rm inc(s)}_{\kappa} \neq 0, \\ a^{\rm inc}_{2\kappa} = a^{\rm inc(w)}_{2\kappa} := 0, \\ a^{\rm inc}_{3\kappa} = a^{\rm inc(w)}_{3\kappa} \neq 0 \end{array}\right\} \quad \Rightarrow \quad \left\{\begin{array}{lll} a^{\rm scat}_{\kappa} = a^{\rm dif}_{\kappa}, & a^{\rm scat}_{2\kappa} = 0, & a^{\rm scat}_{3\kappa} = a^{\rm dif}_{3\kappa} + a^{\rm gen}_{3\kappa} \\ b^{\rm scat}_{\kappa} = b^{\rm dif}_{\kappa}, & b^{\rm scat}_{2\kappa} = 0, & b^{\rm scat}_{3\kappa} = b^{\rm dif}_{3\kappa} + b^{\rm gen}_{3\kappa} \end{array}\right\},$$

$$\left.\begin{array}{l} a^{\rm inc}_{\kappa} = a^{\rm inc(s)}_{\kappa} \neq 0, \\ a^{\rm inc}_{2\kappa} = a^{\rm inc(w)}_{2\kappa} \neq 0, \\ a^{\rm inc}_{3\kappa} = a^{\rm inc(w)}_{3\kappa} := 0 \end{array}\right\} \quad \Rightarrow \quad \left\{\begin{array}{lll} a^{\rm scat}_{\kappa} = a^{\rm dif}_{\kappa} + a^{\rm gen}_{\kappa}, & a^{\rm scat}_{2\kappa} = a^{\rm dif}_{2\kappa}, & a^{\rm scat}_{3\kappa} = a^{\rm gen}_{3\kappa} \\ b^{\rm scat}_{\kappa} = b^{\rm dif}_{\kappa} + b^{\rm gen}_{\kappa}, & b^{\rm scat}_{2\kappa} = b^{\rm dif}_{2\kappa}, & b^{\rm scat}_{3\kappa} = b^{\rm gen}_{3\kappa} \end{array}\right\},$$

$$\left.\begin{array}{l} a^{\rm inc}_{\kappa} = a^{\rm inc(s)}_{\kappa} \neq 0, \\ a^{\rm inc}_{2\kappa} = a^{\rm inc(w)}_{2\kappa} \neq 0, \\ a^{\rm inc}_{3\kappa} = a^{\rm inc(w)}_{3\kappa} \neq 0 \end{array}\right\} \quad \Rightarrow \quad \left\{\begin{array}{lll} a^{\rm scat}_{\kappa} = a^{\rm dif}_{\kappa} + a^{\rm gen}_{\kappa}, & a^{\rm scat}_{2\kappa} = a^{\rm dif}_{2\kappa}, & a^{\rm scat}_{3\kappa} = a^{\rm dif}_{3\kappa} + a^{\rm gen}_{3\kappa} \\ b^{\rm scat}_{\kappa} = b^{\rm dif}_{\kappa} + b^{\rm gen}_{\kappa}, & b^{\rm scat}_{2\kappa} = b^{\rm dif}_{2\kappa}, & b^{\rm scat}_{3\kappa} = b^{\rm dif}_{3\kappa} + b^{\rm gen}_{3\kappa} \end{array}\right\}.$$

The boundary conditions follow from the continuity of the tangential components of the full fields of diffraction $\left\{\mathbf{E}_{\rm tg}(n\kappa;y,z)\right\}_{n=1,2,3}$ $\left\{\mathbf{H}_{\rm tg}(n\kappa;y,z)\right\}_{n=1,2,3}$ at the boundary $z = 2\pi\delta$ and $z = -2\pi\delta$ of the non-linear layer (cf. (C3)). According to (C3) and the presentation of the electrical components of the electromagnetic field (32), at the boundary of the non-linear layer we obtain:

$$\begin{array}{rclcrcl} U(n\kappa;2\pi\delta) & = & a^{\rm scat}_{n\kappa} + a^{\rm inc}_{n\kappa}, & & U'(n\kappa;2\pi\delta) & = & i\Gamma_{n\kappa}\left(a^{\rm scat}_{n\kappa} - a^{\rm inc}_{n\kappa}\right), \\ U(n\kappa;-2\pi\delta) & = & b^{\rm scat}_{n\kappa}, & & U'(n\kappa;-2\pi\delta) & = & -i\Gamma_{n\kappa}b^{\rm scat}_{n\kappa}, \quad n = 1,2,3, \end{array} \tag{33}$$

where "$'$" denotes the differentiation w.r.t. z. Eliminating in (33) the unknown values of the complex amplitudes $\left\{a^{\rm scat}_{n\kappa}\right\}_{n=1,2,3}$, $\left\{b^{\rm scat}_{n\kappa}\right\}_{n=1,2,3}$ of the scattered field and taking into consideration that $a^{\rm inc}_{n\kappa} = U^{\rm inc}(n\kappa;2\pi\delta)$, we arrive at the desired boundary conditions for the problem (31), (C1) – (C4):

$$\begin{array}{rcrcl} i\Gamma_{n\kappa}U(n\kappa;-2\pi\delta) & + & U'(n\kappa;-2\pi\delta) & = & 0, \\ i\Gamma_{n\kappa}U(n\kappa;2\pi\delta) & - & U'(n\kappa;2\pi\delta) & = & 2i\Gamma_{n\kappa}a^{\rm inc}_{n\kappa}, \quad n = 1,2,3. \end{array} \tag{34}$$

Substituting the representation (32) for the desired solution into the system (31), the resulting system of non-linear ordinary differential equations together with the boundary conditions (34) forms a semi-linear boundary-value problem of Sturm-Liouville type, see also Shestopalov & Yatsyk (2010); Yatsyk (2007); Shestopalov & Yatsyk (2007); Angermann & Yatsyk (2010).

5. The system of non-linear integral equations

Similarly to the results given in Yatsyk (2007); Shestopalov & Yatsyk (2007); Kravchenko & Yatsyk (2007); Angermann & Yatsyk (2010); Shestopalov & Sirenko (1989), the problem (31), (C1) – (C4) reduces to finding solutions of one-dimensional non-linear integral equations (along the height $z \in (-2\pi\delta, 2\pi\delta)$ of the structure) w.r.t. the components $U(n\kappa;z)$, $n = 1,2,3$, of the fields scattered and generated in the non-linear layer. We give the derivation of this system of equations in the case of excitation of the non-linear structure by a plane wave packet (30).
The solution of (31), (C1) – (C4) in the whole space $Q := \{q = (y,z) : |y| < \infty, |z| < \infty\}$ is obtained using the properties of the canonical Green's function of the problem (31), (C1) – (C4) (for the special case $\varepsilon_{n\kappa} \equiv 1$) which is defined, for $Y > 0$, in the strip $Q_{\{Y,\infty\}} := \{q = (y,z) : |y| < Y, |z| < \infty\} \subset Q$ by

$$\begin{aligned} & G_0(n\kappa;q,q_0) \\ := \; & \frac{i}{4Y} \exp\left\{i\left[\phi_{n\kappa}(y-y_0) + \Gamma_{n\kappa}|z-z_0|\right]\right\} / \Gamma_{n\kappa} \\ = \; & \exp(\pm i\phi_{n\kappa} y)\, \frac{i\pi}{4Y} \int_{-\infty}^{\infty} H_0^{(1)}\left(n\kappa\sqrt{(\check{y}-y_0)^2 + (z-z_0)^2}\right) \exp(\mp i\phi_{n\kappa}\check{y})\, d\check{y}, \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n = 1,2,3 \end{aligned} \tag{35}$$

(cf. Shestopalov & Sirenko (1989); Sirenko et al. (1985)).
We derive the system of non-linear integral equations by the same classical approach as described in Smirnov (1981) (see also Shestopalov & Yatsyk (2007)). Denote both the scattered and the generated full fields of diffraction at each frequency $n\kappa$, $n = 1,2,3$, i.e. the solution of the problem (31), (C1) – (C4), by $E_1\left(n\kappa;q|_{q=(y,z)}\right) = U(n\kappa;z)\exp(i\phi_{n\kappa}y)$ (cf. (32)), and write the system of equations (31) in the form

$$\begin{cases} \left(\nabla^2 + \kappa^2\right) E_1(\kappa;q) = [1 - \varepsilon_\kappa(q,\alpha(q),E_1(\kappa;q),E_1(2\kappa;q),E_1(3\kappa;q))]\,\kappa^2 E_1(\kappa;q) \\ \qquad\qquad - \alpha(q)\kappa^2 E_1^2(2\kappa;q)E_1^*(3\kappa;q), \\ \left(\nabla^2 + (2\kappa)^2\right) E_1(2\kappa;q) = [1 - \varepsilon_{2\kappa}(q,\alpha(q),E_1(\kappa;q),E_1(2\kappa;q),E_1(3\kappa;q))]\,(2\kappa)^2 E_1(2\kappa;q), \\ \left(\nabla^2 + (3\kappa)^2\right) E_1(3\kappa;q) = [1 - \varepsilon_{3\kappa}(q,\alpha(q),E_1(\kappa;q),E_1(2\kappa;q),E_1(3\kappa;q))]\,(3\kappa)^2 E_1(3\kappa;q) \\ \qquad\qquad - \alpha(q)(3\kappa)^2\left\{\frac{1}{3}E_1^3(\kappa;q) + E_1^2(2\kappa;q)E_1^*(\kappa;q)\right\}, \end{cases}$$

or, shorter,

$$\begin{aligned} \left(\nabla^2 + (n\kappa)^2\right) E_1(n\kappa;q) = \; & [1 - \varepsilon_{n\kappa}(q,\alpha(q),E_1(\kappa;q),E_1(2\kappa;q),E_1(3\kappa;q))]\,(n\kappa)^2 E_1(n\kappa;q) \\ & - \delta_{n1}\alpha(q)(n\kappa)^2 E_1^2(2\kappa;q)E_1^*(3\kappa;q) \\ & - \delta_{n3}\alpha(q)(n\kappa)^2\left\{\frac{1}{3}E_1^3(\kappa;q) + E_1^2(2\kappa;q)E_1^*(\kappa;q)\right\}, \quad n = 1,2,3. \end{aligned} \tag{36}$$

At the right-hand side of the system of equations (36), the first term outside the layer vanishes, since, by assumption, the permittivity of the medium in which the non-linear layer is situated is equal to one, i.e. $1-\varepsilon_{n\kappa}(q,\alpha(q),E_1(\kappa;q),E_1(2\kappa;q),E_1(3\kappa;q))\equiv 0$ for $|z|>2\pi\delta$.
The excitation field of the non-linear structure can be represented in the form of a packet of incident plane waves $\left\{E_1^{\mathrm{inc}}(n\kappa;q)\right\}_{n=1,2,3}$ satisfying the condition of phase synchronism, where

$$E_1^{\mathrm{inc}}(n\kappa;q)=a_{n\kappa}^{\mathrm{inc}}\exp\left\{i\left[\phi_{n\kappa}y-\Gamma_{n\kappa}(z-2\pi\delta)\right]\right\},\quad n=1,2,3. \tag{37}$$

Furthermore, in the present situation described by the system of equations (36), we assume that the excitation field $E_1^{\mathrm{inc}}(\kappa;q)$ of the non-linear structure at the frequency κ is sufficiently strong (i.e. the amplitude a_κ^{inc} is sufficiently large such that the third harmonic generation is possible), whereas the amplitudes $a_{2\kappa}^{\mathrm{inc}}$, $a_{3\kappa}^{\mathrm{inc}}$ corresponding to excitation fields $E_1^{\mathrm{inc}}(2\kappa;q)$, $E_1^{\mathrm{inc}}(3\kappa;q)$ at the frequencies 2κ, 3κ, respectively, are selected sufficiently weak such that no generation of multiple harmonics occurs.
In the whole space, for each frequency $n\kappa$, $n=1,2,3$, the fields $\left\{E_1^{\mathrm{inc}}(n\kappa;q)\right\}_{n=1,2,3}$ of incident plane waves satisfy a system of homogeneous Helmholtz equations:

$$\left(\nabla^2+(n\kappa)^2\right)E_1^{\mathrm{inc}}(n\kappa;q)=0,\quad q\in Q,\quad n=1,2,3. \tag{38}$$

For $z>2\pi\delta$, the incident fields $\left\{E_1^{\mathrm{inc}}(n\kappa;q)\right\}_{n=1,2,3}$ are fields of plane waves approaching the layer, while, for $z<2\pi\delta$, they move away from the layer and satisfy the radiation condition (since, in the representation of the fields $E_1^{\mathrm{inc}}(n\kappa;q)$, $n=1,2,3$, the transverse propagation constants $\Gamma_{n\kappa}>0$, $n=1,2,3$ are positive).
Subtracting the incident fields $E_1^{\mathrm{inc}}(n\kappa;q)$, from the corresponding total fields $E_1(n\kappa;q)$, cf. (32), we obtain the following equations w.r.t. the scattered fields $E_1(n\kappa;q)-E_1^{\mathrm{inc}}(n\kappa;q)=:E_1^{\mathrm{scat}}(n\kappa;q)$ in the zone of reflection $z>2\pi\delta$, the fields $E_1(n\kappa;q)$, $|z|\le 2\pi\delta$, scattered in the layer and the fields $E_1(n\kappa;q)=:E_1^{\mathrm{scat}}(n\kappa;q)$, $z<2\pi\delta$, passing through the layer:

$$\begin{aligned}
\left(\nabla^2+(n\kappa)^2\right)\left[E_1(n\kappa;q)\right. &- \left.E_1^{\mathrm{inc}}(n\kappa;q)\right] = 0,\quad z>2\pi\delta,\\
\left(\nabla^2+(n\kappa)^2\right)E_1(n\kappa;q) &= \left[1-\varepsilon_{n\kappa}(q,\alpha(q),E_1(\kappa;q),E_1(2\kappa;q),E_1(3\kappa;q))\right](n\kappa)^2E_1(n\kappa;q)\\
&\quad -\delta_{n1}\alpha(q)(n\kappa)^2E_1^2(2\kappa;q)E_1^*(3\kappa;q)\\
&\quad -\delta_{n3}\alpha(q)(n\kappa)^2\left\{\tfrac{1}{3}E_1^3(\kappa;q)+E_1^2(2\kappa;q)E_1^*(\kappa;q)\right\},\quad |z|\le 2\pi\delta,\\
\left(\nabla^2+(n\kappa)^2\right)E_1(n\kappa;q) &= 0,\quad z<-2\pi\delta,\quad n=1,2,3.
\end{aligned} \tag{39}$$

Since the canonical Green's functions satisfy the equations

$$\left(\nabla^2+(n\kappa)^2\right)G_0(n\kappa;q,q_0)=-\delta(q,q_0),\quad n=1,2,3, \tag{40}$$

where $\delta(q,q_0)$ denotes the Dirac delta-function, it is easy to obtain from the above equations (39), with q replaced by q_0, the following system:

$$
\begin{aligned}
& \left[E_1(n\kappa;q_0)-E_1^{\rm inc}(n\kappa;q_0)\right]\nabla^2 G_0(n\kappa;q,q_0)-G_0(n\kappa;q,q_0)\nabla^2\left[E_1(n\kappa;q_0)-E_1^{\rm inc}(n\kappa;q_0)\right] \\
= \;& -\left[E_1(n\kappa;q_0)-E_1^{\rm inc}(n\kappa;q_0)\right]\delta(q,q_0), \quad z>2\pi\delta, \\
& E_1(n\kappa;q_0)\nabla^2 G_0(n\kappa;q,q_0)-G_0(n\kappa;q,q_0)\nabla^2 E_1(n\kappa;q_0) \\
= \;& -E_1(n\kappa;q_0)\delta(q,q_0) \\
& -G_0(n\kappa;q,q_0)\left[1-\varepsilon_{n\kappa}\left(q_0,\alpha(q_0),E_1(\kappa;q_0),E_1(2\kappa;q_0),E_1(3\kappa;q_0)\right)\right](n\kappa)^2 E_1(n\kappa;q_0) \\
& +\delta_{n1}G_0(n\kappa;q,q_0)\alpha(q)(n\kappa)^2 E_1^2(2\kappa;q)E_1^*(3\kappa;q) \\
& +\delta_{n3}G_0(n\kappa;q,q_0)\alpha(q)(n\kappa)^2\left\{\tfrac{1}{3}E_1^3(\kappa;q)+E_1^2(2\kappa;q)E_1^*(\kappa;q)\right\}, \quad |z|\le 2\pi\delta, \\
& E_1(n\kappa;q_0)\nabla^2 G_0(n\kappa;q,q_0)-G_0(n\kappa;q,q_0)\nabla^2 E_1(n\kappa;q_0) \\
= \;& -E_1(n\kappa;q_0)\delta(q,q_0), \quad z<-2\pi\delta, \quad n=1,2,3.
\end{aligned}
\tag{41}
$$

Given $Y>0$, $Z>2\pi\delta$, now we consider in the space Q the rectangular domain

$$Q_{\{Y,Z\}}:=\{q=(y,z):\ |y|<Y,\ |z|<Z\},$$

and the subsets

$$
\begin{aligned}
Q_{\{Y,Z\},\,z>2\pi\delta} &:= \{q=(y,z):\ |y|<Y,\ 2\pi\delta<z\le Z\}, \\
Q_{\{Y,Z\},\,|z|\le 2\pi\delta} &:= \{q=(y,z):\ |y|<Y,\ |z|\le 2\pi\delta\}, \\
Q_{\{Y,Z\},\,z<-2\pi\delta} &:= \{q=(y,z):\ |y|<Y,\ -Z\le z<-2\pi\delta\},
\end{aligned}
$$

and make use of Green's formula.

We also mention that in the case of a non-linear layered structure consisting of a finite number of layers the applicability of Green's formula in the region $Q_{\{Y,Z\},\,|z|\le 2\pi\delta}$ occupied by the dielectric follows from the continuity condition (C3) w.r.t. $\mathbf{E}_{\rm tg}(n\kappa;q)$, $\mathbf{H}_{\rm tg}(n\kappa;q)$ at the boundaries. Indeed, consider a covering of $Q_{\{Y,Z\}}$ by a finite number of disjoint rectangles such that the restrictions of $\varepsilon_{n\kappa}\left(q_0,\alpha(q_0),E_1(\kappa;q_0),E_1(2\kappa;q_0),E_1(3\kappa;q_0)\right)$ to each of these rectangles are smooth functions. At the common interfaces of these regions (i.e. at the boundaries of the separate layers of the structure) due to the continuity of the components $\mathbf{E}_{\rm tg}(n\kappa;q)$ and $\mathbf{H}_{\rm tg}(n\kappa;q)$ of the electromagnetic field (cf. (C3)), $E_1(n\kappa;q)$ and $\partial E_1(n\kappa;q)/\partial\mathbf{n}$ are continuous (where $\mathbf{n}$ denotes the outward unit normal w.r.t. each of the regions). Now, by Green's formula and condition (C3) it is easy to obtain the system of non-linear integral equations w.r.t. the unknown solutions $E_1(n\kappa;q)$, $n=1,2,3$, in the region $Q_{\{Y,Z\},\,|z|\le 2\pi\delta}$. This system forms an integral representation of the solution in the exterior $Q_{\{Y,Z\}}\setminus Q_{\{Y,Z\},\,|z|\le 2\pi\delta}$ of the region occupied by the dielectric layer. Consequently, the desired functions $\{E_1(n\kappa;q)\}_{n=1,2,3}$, which are twice continuously differentiable both within (i.e. $Q_{\{Y,Z\},\,|z|\le 2\pi\delta}$) and outside (i.e. $Q_{\{Y,Z\},\,|z|>2\pi\delta}$) of the region occupied by the dielectric layer, are continuous and have continuous derivatives throughout the whole region $Q_{\{Y,Z\}}$ up to and including the boundary $\partial Q_{\{Y,Z\}}$, i.e. $E_1(n\kappa;q)\in C^2\left(Q_{\{Y,Z\}}\right)\cap C^1\left(\overline{Q_{\{Y,Z\}}}\right)$, $n=1,2,3$. The system of non-linear integral equations and the corresponding integral representations of the desired solution are obtained by applying, in each of the rectangles $Q_{\{Y,Z\},\,z>2\pi\delta}$, $Q_{\{Y,Z\},\,|z|\le 2\pi\delta}$, $Q_{\{Y,Z\},\,z<-2\pi\delta}$, Green's formula to the functions $E_1(n\kappa;q_0)-E_1^{\rm inc}(n\kappa;q_0)=:E_1^{\rm scat}(n\kappa;q_0)$ for $q_0\in Q_{\{Y,Z\},\,z>2\pi\delta}$, $E_1(n\kappa;q_0)=:E_1^{\rm scat}(n\kappa;q_0)$ for $q_0\in Q_{\{Y,Z\},\,|z|\le 2\pi\delta}$, $E_1(n\kappa;q_0)=:E_1^{\rm scat}(n\kappa;q_0)$ for $q_0\in Q_{\{Y,Z\}}$, $z<-2\pi\delta$, and $G_0(n\kappa;q,q_0)$ for $q,q_0\in Q_{\{Y,Z\}}$:

$$\iint_{Q_{\{Y,Z\},\,z>2\pi\delta}} \left(\left[E_1 - E_1^{\text{inc}}\right] \nabla^2 G_0 - G_0 \nabla^2 \left[E_1 - E_1^{\text{inc}}\right] \right) dq_0$$

$$= \int\limits_{Q_{\{Y,Z\},\,z>2\pi\delta}} \left(\left[E_1 - E_1^{\text{inc}}\right] \frac{\partial G_0}{\partial \mathbf{n}} - G_0 \frac{\partial \left[E_1 - E_1^{\text{inc}}\right]}{\partial \mathbf{n}} \right) dq_0,$$

$$\iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} \left(E_1 \nabla^2 G_0 - G_0 \nabla^2 E_1 \right) dq_0 = \int_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} \left(E_1 \frac{\partial G_0}{\partial \mathbf{n}} - G_0 \frac{\partial E_1}{\partial \mathbf{n}} \right) dq_0,$$

$$\iint_{Q_{\{Y,Z\},\,z<-2\pi\delta}} \left(E_1 \nabla^2 G_0 - G_0 \nabla^2 E_1 \right) dq_0 = \int_{Q_{\{Y,Z\},\,z<-2\pi\delta}} \left(E_1 \frac{\partial G_0}{\partial \mathbf{n}} - G_0 \frac{\partial E_1}{\partial \mathbf{n}} \right) dq_0, \quad n = 1,2,3. \tag{42}$$

Taking into account the relations (41), we get

$$\begin{cases} E_1(n\kappa;q) - E_1^{\text{inc}}(n\kappa;q), & q \in Q_{\{Y,Z\},\,z>2\pi\delta} \\ 0, & q \in Q_{\{Y,Z\}} \setminus \partial Q_{\{Y,Z\},\,z>2\pi\delta} \end{cases}$$

$$= -\int_{\partial Q_{\{Y,Z\},\,z>2\pi\delta}} \left(\left[E_1(n\kappa;q_0) - E_1^{\text{inc}}(n\kappa;q_0)\right] \frac{\partial G_0(n\kappa;q,q_0)}{\partial \mathbf{n}} - G_0(n\kappa;q,q_0) \frac{\partial \left[E_1(n\kappa;q_0) - E_1^{\text{inc}}(n\kappa;q_0)\right]}{\partial \mathbf{n}} \right) dq_0,$$

$$\begin{cases} E_1(n\kappa;q), & q \in Q_{\{Y,Z\},\,|z|\le 2\pi\delta} \\ 0, & q \in Q_{\{Y,Z\}} \setminus Q_{\{Y,Z\},\,|z|\le 2\pi\delta} \end{cases}$$

$$\begin{aligned} = & -(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0) \times \\ & \times \left[1 - \varepsilon_{n\kappa}(q_0, \alpha(q_0), E_1(\kappa;q_0), E_1(2\kappa;q_0), E_1(3\kappa;q_0))\right] E_1(n\kappa;q_0) dq_0 \\ & + \delta_{n1}(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0) \alpha(q_0) E_1^2(2\kappa;q_0) E_1^*(3\kappa;q_0) dq_0 \\ & + \delta_{n3}(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0) \alpha(q_0) \left\{ \frac{1}{3} E_1^3(\kappa;q_0) + E_1^2(2\kappa;q_0) E_1^*(\kappa;q_0) \right\} dq_0 \\ & - \int_{\partial Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} \left(E_1(n\kappa;q_0) \frac{\partial G_0(n\kappa;q,q_0)}{\partial \mathbf{n}} - G_0(n\kappa;q,q_0) \frac{\partial E_1(n\kappa;q_0)}{\partial \mathbf{n}} \right) dq_0, \end{aligned}$$

$$\begin{cases} E_1(n\kappa;q), & q \in Q_{\{Y,Z\},\,z<-2\pi\delta} \\ 0, & q \in Q_{\{Y,Z\}} Q_{\{Y,Z\},\,z<-2\pi\delta} \end{cases}$$

$$= -\int_{\partial Q_{\{Y,Z\},\,z<-2\pi\delta}} \left(E_1(n\kappa;q_0) \frac{\partial G_0(n\kappa;q,q_0)}{\partial \mathbf{n}} - G_0(n\kappa;q,q_0) \frac{\partial E_1(n\kappa;q_0)}{\partial \mathbf{n}} \right) dq_0, \quad n = 1,2,3. \tag{43}$$

Suppose $q \in Q_{\{Y,Z\},\,|z|\le 2\pi\delta}$, i.e. it lies in a rectangle containing the non-linear structure. Then the equations of (43) take the form

$$
\begin{aligned}
0 &= -\int_{\partial Q_{\{Y,Z\},\, z>2\pi\delta}} \left(\left[E_1(n\kappa;q_0) - E_1^{\mathrm{inc}}(n\kappa;q_0) \right] \frac{\partial G_0(n\kappa;q,q_0)}{\partial \mathbf{n}} \right. \\
&\qquad \left. - G_0(n\kappa;q,q_0) \frac{\partial \left[E_1(n\kappa;q_0) - E_1^{\mathrm{inc}}(n\kappa;q_0) \right]}{\partial \mathbf{n}} \right) dq_0, \\
E_1(n\kappa;q) &= -(n\kappa)^2 \iint_{Q_{\{Y,Z\},\, |z|\le 2\pi\delta}} G_0(n\kappa;q,q_0) \times \\
&\qquad \times \left[1 - \varepsilon_{n\kappa}\left(q_0, \alpha(q_0), E_1(\kappa;q_0), E_1(2\kappa;q_0), E_1(3\kappa;q_0)\right) \right] E_1(n\kappa;q_0) dq_0 \\
&\quad + \delta_{n1}(n\kappa)^2 \iint_{Q_{\{Y,Z\},\, |z|\le 2\pi\delta}} G_0(n\kappa;q,q_0)\alpha(q_0) E_1^2(2\kappa;q_0) E_1^*(3\kappa;q_0) dq_0 \\
&\quad + \delta_{n3}(n\kappa)^2 \iint_{Q_{\{Y,Z\},\, |z|\le 2\pi\delta}} G_0(n\kappa;q,q_0)\alpha(q_0) \times \\
&\qquad \times \left\{ \frac{1}{3} E_1^3(\kappa;q_0) + E_1^2(2\kappa;q_0) E_1^*(\kappa;q_0) \right\} dq_0 \\
&\quad - \int_{\partial Q_{\{Y,Z\},\, |z|\le 2\pi\delta}} \left(E_1(n\kappa;q_0) \frac{\partial G_0(n\kappa;q,q_0)}{\partial \mathbf{n}} - G_0(n\kappa;q,q_0) \frac{\partial E_1(n\kappa;q_0)}{\partial \mathbf{n}} \right) dq_0, \\
0 &= -\int_{\partial Q_{\{Y,Z\},\, z<-2\pi\delta}} \left(E_1(n\kappa;q_0) \frac{\partial G_0(n\kappa;q,q_0)}{\partial \mathbf{n}} - G_0(n\kappa;q,q_0) \frac{\partial E_1(n\kappa;q_0)}{\partial \mathbf{n}} \right) dq_0, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n = 1,2,3.
\end{aligned}
\tag{44}
$$

If the parameter Z increases to infinity, $Z \to \infty$, the line integrals appearing in the first and third equations of (44) along the lower $[(-Z,-Y),(-Z,Y)]$ and upper $[(Z,Y),(Z,-Y)]$ parts of the boundary $\partial Q_{\{Y,Z\}}$ tend to zero for all $n = 1,2,3$. This is a consequence of the fact that, for all frequencies $n\kappa$, $n = 1,2,3$, the reflected field $E_1(n\kappa;q) - E_1^{\mathrm{inc}}(n\kappa;q) =: E_1^{\mathrm{scat}}(n\kappa;q)$, given by the first equation of (44), and the field $E_1(n\kappa;q) =: E_1^{\mathrm{scat}}(n\kappa;q)$, passing through the layer and described by the third equation of (44), satisfy the radiation condition (C4), and of the asymptotic properties of the canonical Green's function (35). The line integrals along the left $[(-Z,Y),\ (Z,Y)]$ and right $[(Z,-Y),\ (-Z,-Y)]$ sides of the boundary $\partial Q_{\{Y,Z\}}$ cancel out each other in all equations of the system (44).

Next we consider the components of the total fields $E_1(n\kappa;q)$ (i.e. $\mathbf{E}_{\mathrm{tg}}(n\kappa;q)$ and $\frac{\partial E_1(n\kappa;q)}{\partial \mathbf{n}}$) (i.e. $\mathbf{H}_{\mathrm{tg}}(n\kappa;q)$) at the common boundaries of neighbouring rectangles. At the upper $z = 2\pi\delta$ and lower $z = -2\pi\delta$ boundaries of the non-linear medium, they are continuous, cf. the interface condition (C3). The orientations of the outer normals in the line integrals of the system (44) (for the first and second equations, and for the second and third equations, for each $n = 1,2,3$) at these common boundaries are opposite. Adding all equations of the system (44), we obtain

$$
\begin{aligned}
& E_1(n\kappa;q) \\
= \; & -(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0) \times \\
& \qquad \times \left[1 - \varepsilon_{n\kappa}\left(q_0, \alpha(q_0), E_1(\kappa;q_0), E_1(2\kappa;q_0), E_1(3\kappa;q_0)\right)\right] E_1(n\kappa;q_0)\,dq_0 \\
& + \delta_{n1}(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0)\alpha(q_0)E_1^2(2\kappa;q_0)\,E_1^*(3\kappa;q_0)\,dq_0 \\
& + \delta_{n3}(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0)\alpha(q_0)\left\{\frac{1}{3}E_1^3(\kappa;q_0) + E_1^2(2\kappa;q_0)E_1^*(\kappa;q_0)\right\} dq_0 \\
& + \int_{\underline{\partial Q_{\{Y,Z=\infty\},\,z>2\pi\delta}}} \left(E_1(n\kappa;q_0)\frac{\partial G_0(n\kappa;q,q_0)}{\partial \mathbf{n}} - G_0(n\kappa;q,q_0)\frac{\partial E_1(n\kappa;q_0)}{\partial \mathbf{n}}\right) dq_0, \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad q \in Q_{\{Y,Z\},\,|z|\le 2\pi\delta}, \quad n = 1,2,3.
\end{aligned}
\tag{45}
$$

In the line integrals of equation (45), at each of the frequencies $n\kappa$, $n = 1,2,3$, the integration runs along the lower boundary $\underline{\partial Q_{\{Y,Z=\infty\},\,z>2\pi\delta}}$ of the half-space $Q_{\{Y,Z=\infty\},\,z>2\pi\delta}$, where the normal vector $\mathbf{n}$ points into the non-linear layer. Changing the orientation of the normal vector (which is equivalent to changing the sign of the integral) and considering the line integrals as integrals along the upper boundary $\overline{\partial Q_{\{Y,Z\},\,z\le 2\pi\delta}}$ of the region $Q_{\{Y,Z\},\,z\le 2\pi\delta}$, we get

$$
\begin{aligned}
& E_1(n\kappa;q) \\
= \; & -(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0) \times \\
& \qquad \times \left[1 - \varepsilon_{n\kappa}\left(q_0, \alpha(q_0), E_1(\kappa;q_0), E_1(2\kappa;q_0), E_1(3\kappa;q_0)\right)\right] E_1(n\kappa;q_0)\,dq_0 \\
& + \delta_{n1}(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0)\alpha(q_0)E_1^2(2\kappa;q_0)E_1^*(3\kappa;q_0)\,dq_0 \\
& + \delta_{n3}(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0)\alpha(q_0)\left\{\frac{1}{3}E_1^3(\kappa;q_0) + E_1^2(2\kappa;q_0)E_1^*(\kappa;q_0)\right\} dq_0 \\
& - \int_{\overline{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}}} \left(E_1(n\kappa;q_0)\frac{\partial G_0(n\kappa;q,q_0)}{\partial \mathbf{n}} - G_0(n\kappa;q,q_0)\frac{\partial E_1(n\kappa;q_0)}{\partial \mathbf{n}}\right) dq_0, \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad q \in Q_{\{Y,Z\},\,|z|\le 2\pi\delta}, \quad n = 1,2,3.
\end{aligned}
\tag{46}
$$

The line integrals in (46) represent the values of the incident fields at the frequencies $n\kappa$, $n = 1,2,3$, in the points $q \in Q_{\{Y,Z\},\,|z|\le 2\pi\delta}$:

$$
\begin{aligned}
E_1^{\mathrm{inc}}(n\kappa;q) \quad = \quad & -\int_{\overline{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}}} \left(E_1(n\kappa;q_0)\frac{\partial G_0(n\kappa;q,q_0)}{\partial \mathbf{n}} - G_0(n\kappa;q,q_0)\frac{\partial E_1(n\kappa;q_0)}{\partial \mathbf{n}}\right) dq_0, \\
& \qquad\qquad\qquad\qquad\qquad\qquad q \in Q_{\{Y,Z\},\,|z|\le 2\pi\delta}, \quad n = 1,2,3.
\end{aligned}
\tag{47}
$$

Indeed, applying Green's formula to the functions $G(n\kappa;q,q_0)$ and $E_1^{\mathrm{inc}}(n\kappa;q)$ in the region $Q_{\{Y,Z\},\,|z|\le 2\pi\delta} \cup Q_{\{Y,Z\},\,z<-2\pi\delta}$ (where $q \in Q_{\{Y,Z\},\,|z|\le 2\pi\delta} \cup Q_{\{Y,Z\},\,z<-2\pi\delta}$) and letting $\partial Q_{\{Y,Z\},\,z<-2\pi\delta} \to -\infty$, we arrive at (47). Substituting (47) into (46), we obtain the following system of non-linear integral equations w.r.t. the unknown total diffraction field:

$$
\begin{aligned}
& E_1(n\kappa;q) \\
= \; & -(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0) \times \\
& \qquad \times \left[1-\varepsilon_{n\kappa}\left(q_0,\alpha(q_0),E_1(\kappa;q_0),E_1(2\kappa;q_0),E_1(3\kappa;q_0)\right)\right] E_1(n\kappa;q_0)dq_0 \\
& + \delta_{n1}(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0)\alpha(q_0)E_1^2(2\kappa;q_0)E_1^*(3\kappa;q_0)dq_0 \\
& + \delta_{n3}(n\kappa)^2 \iint_{Q_{\{Y,Z\},\,|z|\le 2\pi\delta}} G_0(n\kappa;q,q_0)\alpha(q_0)\left\{\frac{1}{3}E_1^3(\kappa;q_0)+E_1^2(2\kappa;q_0)E_1^*(\kappa;q_0)\right\}dq_0 \\
& + E_1^{\mathrm{inc}}(n\kappa;q), \quad q\in Q_{\{Y,Z\},\,|z|\le 2\pi\delta}, \quad n=1,2,3.
\end{aligned}
$$

Passing in the above equations to the limit $Y\to\infty$ (where this procedure is admissible because of the free choice of the parameter Y and the asymptotic behaviour of the integrands as $\mathcal{O}\left(Y^{-1}\right)$, see (C1) and (35)) we arrive at a system of non-linear integral equations w.r.t. the total diffraction fields in the strip $Q_\delta := Q_{\{Y=\infty,Z\},\,|z|\le 2\pi\delta} = \{q=(y,z):\ |y|<\infty,\ |z|\le 2\pi\delta\}$ filled by the non-linear dielectric layer:

$$
\begin{aligned}
& E_1(n\kappa;q) \\
= \; & -(n\kappa)^2 \iint_{Q_\delta} G_0(n\kappa;q,q_0) \times \\
& \qquad \times \left[1-\varepsilon_{n\kappa}\left(q_0,\alpha(q_0),E_1(\kappa;q_0),E_1(2\kappa;q_0),E_1(3\kappa;q_0)\right)\right] E_1(n\kappa;q_0)dq_0 \\
& + \delta_{n1}(n\kappa)^2 \iint_{Q_\delta} G_0(n\kappa;q,q_0)\alpha(q_0)E_1^2(2\kappa;q_0)E_1^*(3\kappa;q_0)dq_0 \\
& + \delta_{n3}(n\kappa)^2 \iint_{Q_\delta} G_0(n\kappa;q,q_0)\alpha(q_0)\left\{\frac{1}{3}E_1^3(\kappa;q_0)+E_1^2(2\kappa;q_0)E_1^*(\kappa;q_0)\right\}dq_0 \\
& + E_1^{\mathrm{inc}}(n\kappa;q), \quad q\in Q_\delta, \quad n=1,2,3.
\end{aligned}
\tag{48}
$$

The integral representations of the total diffraction fields $E_1(n\kappa;q)$, $n=1,2,3$, in the points $q\notin Q_\delta$ located outside the layer can be derived similarly to the approach described above (see (35) – (48)). For this situation it is sufficient to consider in (43) the points lying above ($q\in Q_{\{Y=\infty,Z=\infty\},\,z>2\pi\delta}$) and below ($q\in Q_{\{Y=\infty,Z=\infty\},\,z<-2\pi\delta}$) the layer. As a result, we get that the integral representations (48) are valid for all points in the region $q\in Q := Q_{\{Y=\infty,Z=\infty\},\,z>2\pi\delta}\cup Q_\delta\cup Q_{\{Y=\infty,Z=\infty\},\,z<-2\pi\delta}$, that is

$$
\begin{aligned}
& E_1(n\kappa;q) \\
= \; & -(n\kappa)^2 \iint_{Q_\delta} G_0(n\kappa;q,q_0) \times \\
& \qquad \times \left[1-\varepsilon_{n\kappa}\left(q_0,\alpha(q_0),E_1(\kappa;q_0),E_1(2\kappa;q_0),E_1(3\kappa;q_0)\right)\right] E_1(n\kappa;q_0)dq_0 \\
& + \delta_{n1}(n\kappa)^2 \iint_{Q_\delta} G_0(n\kappa;q,q_0)\alpha(q_0)E_1^2(2\kappa;q_0)E_1^*(3\kappa;q_0)dq_0 \\
& + \delta_{n3}(n\kappa)^2 \iint_{Q_\delta} G_0(n\kappa;q,q_0)\alpha(q_0)\left\{\frac{1}{3}E_1^3(\kappa;q_0)+E_1^2(2\kappa;q_0)E_1^*(\kappa;q_0)\right\}dq_0 \\
& + E_1^{\mathrm{inc}}(n\kappa;q), \quad q\in Q, \quad n=1,2,3.
\end{aligned}
\tag{49}
$$

The expressions in (49) form a system of non-linear integral equations in the points $q\in Q_\delta$. Provided that a solution of this system exists, it can be substituted into the right-hand side of (49). In this way we also obtain an integral representation of the total diffraction field at points located outside the layer, i.e. $q\in Q_{\{Y=\infty,Z=\infty\},\,z>2\pi\delta}$ or $q\in Q_{\{Y=\infty,Z=\infty\},\,z<-2\pi\delta}$.
Alternatively, the system (49) can be derived by means of an iterative approach developed in Shestopalov & Sirenko (1989). Schematically it can be represented as follows (see also Yatsyk

(2007)). In the region Q we construct a sequence $\left\{E_{1,p}(n\kappa;q)\right\}_{p=0}^{\infty}$, $n=1,2,3$, of functions (where each function, starting with the index $p=1$, satisfies the conditions (C1) – (C4)) such that the limit functions $E_1(n\kappa;q)=\lim\limits_{p\to\infty}E_{1,p}(n\kappa;q)$ at the frequencies $n\kappa$, $n=1,2,3$, satisfy (31), (C1) – (C4), i.e.

$$
\begin{aligned}
&\left(\nabla^2+(n\kappa)^2\right)E_{1,0}(n\kappa;q)=0,\\
&\left(\nabla^2+(n\kappa)^2\right)E_{1,1}(n\kappa;q)\\
=\;&\left[1-\varepsilon_{n\kappa}\left(q,\alpha(q),E_{1,0}(\kappa;q),E_{1,0}(2\kappa;q),E_{1,0}(3\kappa;q)\right)\right](n\kappa)^2E_{1,0}(n\kappa;q)\\
&-\delta_{n1}\alpha(q)(n\kappa)^2E_{1,0}^2(2\kappa;q)E_{1,0}^*(3\kappa;q)\\
&-\delta_{n3}\alpha(q)(n\kappa)^2\left\{\tfrac{1}{3}E_{1,0}^3(\kappa;q)+E_{1,0}^2(2\kappa;q)E_{1,0}^*(\kappa;q)\right\},\ldots,\\
&\left(\nabla^2+(n\kappa)^2\right)E_{1,p+1}(n\kappa;q)\\
=\;&\left[1-\varepsilon_{n\kappa}\left(q,\alpha(q),E_{1,p}(\kappa;q),E_{1,p}(2\kappa;q),E_{1,p}(3\kappa;q)\right)\right](n\kappa)^2E_{1,p}(n\kappa;q)\\
&-\delta_{n1}\alpha(q)(n\kappa)^2E_{1,p}^2(2\kappa;q)E_{1,p}^*(3\kappa;q)\\
&-\delta_{n3}\alpha(q)(n\kappa)^2\left\{\tfrac{1}{3}E_{1,p}^3(\kappa;q)+E_{1,p}^2(2\kappa;q)E_{1,p}^*(\kappa;q)\right\},\ldots,\\
&\qquad\qquad n=1,2,3.
\end{aligned}
\tag{50}
$$

The system of equations (50) is formally equivalent to the following:

$$
\begin{aligned}
&E_{1,0}(n\kappa;q):=E_1^{\mathrm{inc}}(n\kappa;q),\\
&E_{1,1}(n\kappa;q)\\
=\;&-(n\kappa)^2\iint_{Q_\delta}G_0(n\kappa;q,q_0)\times\\
&\quad\times\left[1-\varepsilon_{n\kappa}\left(q_0,\alpha(q_0),E_{1,0}(\kappa;q_0),E_{1,0}(2\kappa;q_0),E_{1,0}(3\kappa;q_0)\right)\right]E_{1,0}(n\kappa;q_0)dq_0\\
&+\delta_{n1}(n\kappa)^2\iint_{Q_\delta}G_0(n\kappa;q,q_0)\alpha(q_0)E_{1,0}^2(2\kappa;q_0)E_{1,0}^*(3\kappa;q_0)dq_0\\
&+\delta_{n3}(n\kappa)^2\iint_{Q_\delta}G_0(n\kappa;q,q_0)\alpha(q_0)\left\{\frac{1}{3}E_{1,0}^3(\kappa;q_0)+E_{1,0}^2(2\kappa;q_0)E_{1,0}^*(\kappa;q_0)\right\}dq_0\\
&+E_{1,0}(n\kappa;q),\ldots,\\
&E_{1,p+1}(n\kappa;q)\\
=\;&-(n\kappa)^2\iint_{Q_\delta}G_0(n\kappa;q,q_0)\times\\
&\quad\times\left[1-\varepsilon_{n\kappa}\left(q_0,\alpha(q_0),E_{1,p}(\kappa;q_0),E_{1,p}(2\kappa;q_0),E_{1,p}(3\kappa;q_0)\right)\right]E_{1,p}(n\kappa;q_0)dq_0\\
&+\delta_{n1}(n\kappa)^2\iint_{Q_\delta}G_0(n\kappa;q,q_0)\alpha(q_0)E_{1,p}^2(2\kappa;q_0)E_{1,p}^*(3\kappa;q_0)dq_0\\
&+\delta_{n3}(n\kappa)^2\iint_{Q_\delta}G_0(n\kappa;q,q_0)\alpha(q_0)\left\{\frac{1}{3}E_{1,p}^3(\kappa;q_0)+E_{1,p}^2(2\kappa;q_0)E_{1,p}^*(\kappa;q_0)\right\}dq_0\\
&+E_{1,0}(n\kappa;q),\ldots,\\
&\qquad\qquad q\in Q,\quad n=1,2,3.
\end{aligned}
\tag{51}
$$

Letting in (51) p tend to infinity, we obtain (49) – the integral representations of the unknown diffraction fields in the region Q.

We consider now the variation of the parameter q in the strip occupied by the dielectric layer, i.e. $q\in Q_\delta$. Then the representation (49) can be converted into a system of non-linear integral equations w.r.t. the unknown fields $E_1(n\kappa;q)$, $n=1,2,3$, $q\in Q_\delta$, scattered in the non-linear structure, see (32). Namely, substituting the representations for the canonical Green's functions (35) into (49) and taking into consideration the expressions for the permittivity

$$\varepsilon_{n\kappa}(q_0,\alpha(q_0),E_1(\kappa;q_0),E_1(2\kappa;q_0),E_1(3\kappa;q_0)) = \varepsilon_{n\kappa}(z_0,\alpha(z_0),U(\kappa;z_0),U(2\kappa;z_0),U(3\kappa;z_0)),$$

we get the following system w.r.t. the unknown quasi-homogeneous fields
$E_1\left(n\kappa;q|_{q\equiv(y,z)}\right) = U(n\kappa;z)\exp(i\phi_{n\kappa}y)$, $n=1,2,3$, $|z|\le 2\pi\delta$:

$$\begin{aligned}
& U(n\kappa;z)\exp(i\phi_{n\kappa}y) \\
= & -\lim_{Y\to\infty}\left(\frac{i(n\kappa)^2}{4Y\Gamma_{n\kappa}}\exp(i\phi_{n\kappa}y)\int_{-2\pi\delta}^{2\pi\delta}\int_{-Y}^{Y}\exp(i\Gamma_{n\kappa}|z-z_0|)\times\right. \\
& \left.\times\left[1-\varepsilon_{n\kappa}(z_0,\alpha(z_0),U(\kappa;z_0),U(2\kappa;z_0),U(3\kappa;z_0))\right]U(n\kappa;z_0)\,dy_0dz_0\right) \\
& +\lim_{Y\to\infty}\left(\delta_{n1}\frac{i(n\kappa)^2}{4Y\Gamma_{n\kappa}}\exp(i\phi_{n\kappa}y)\times\right. \\
& \left.\times\int_{-2\pi\delta}^{2\pi\delta}\int_{-Y}^{Y}\exp(i\Gamma_{n\kappa}|z-z_0|)\alpha(z_0)U^2(2\kappa;z_0)U^*(3\kappa;z_0)dy_0dz_0\right) \\
& +\lim_{Y\to\infty}\left(\delta_{n3}\frac{i(n\kappa)^2}{4Y\Gamma_{n\kappa}}\exp(i\phi_{n\kappa}y)\times\right. \\
& \left.\times\int_{-2\pi\delta}^{2\pi\delta}\int_{-Y}^{Y}\exp(i\Gamma_{n\kappa}|z-z_0|)\alpha(z_0)\left\{\frac{1}{3}U^3(\kappa;z_0)+U^2(2\kappa;z_0)U^*(\kappa;z_0)\right\}dy_0dz_0\right) \\
& +U^{\rm inc}(n\kappa;z)\exp(i\phi_{n\kappa}y),\quad |z|\le 2\pi\delta,\quad n=1,2,3.
\end{aligned}$$

Integrating in the region Q_δ w.r.t. the variable y_0, we arrive at a system of non-linear Fredholm integral equations of the second kind w.r.t. the unknown functions $U(n\kappa;z)\in L_2(-2\pi\delta,2\pi\delta)$:

$$\begin{aligned}
& U(n\kappa;z)+\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\int_{-2\pi\delta}^{2\pi\delta}\exp(i\Gamma_{n\kappa}|z-z_0|)\times \\
& \qquad\times\left[1-\varepsilon_{n\kappa}(z_0,\,\alpha(z_0),U(\kappa;z_0),U(2\kappa;z_0),U(3\kappa;z_0))\right]U(n\kappa;z_0)dz_0 \\
= & \ \delta_{n1}\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\int_{-2\pi\delta}^{2\pi\delta}\exp(i\Gamma_{n\kappa}|z-z_0|)\alpha(z_0)U^2(2\kappa;z_0)U^*(3\kappa;z_0)dz_0 \\
& +\delta_{n3}\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\int_{-2\pi\delta}^{2\pi\delta}\exp(i\Gamma_{n\kappa}|z-z_0|)\alpha(z_0)\left\{\frac{1}{3}U^3(\kappa;z_0)+U^2(2\kappa;z_0)U^*(\kappa;z_0)\right\}dz_0 \\
& +U^{\rm inc}(n\kappa;z),\quad |z|\le 2\pi\delta,\quad n=1,2,3.
\end{aligned} \tag{52}$$

Here $U^{\rm inc}(n\kappa;z)=a^{\rm inc}_{n\kappa}\exp\left[-i\Gamma_{n\kappa}(z-2\pi\delta)\right]$, $n=1,2,3$.
The solution of the original problem (31), (C1) – (C4), represented as (32), can be obtained from (52) using the formulas

$$U(n\kappa;2\pi\delta)=a^{\rm inc}_{n\kappa}+a^{\rm scat}_{n\kappa},\quad U(n\kappa;-2\pi\delta)=b^{\rm scat}_{n\kappa},\quad n=1,2,3, \tag{53}$$

(cf. (C3)).
The derivation of the system of non-linear integral equations (52) shows that (52) can be regarded as an integral representation of the desired solution of (31), (C1) – (C4) (i.e. solutions of the form $E_1(n\kappa;y,z)=U(n\kappa;z)\exp(i\phi_{n\kappa}y)$, $n=1,2,3$, see (32)) for points located outside the non-linear layer: $\{(y,z):\ |y|<\infty,\ |z|>2\pi\delta\}$. Indeed, given the solution of non-linear integral equations (52) in the region $|z|\le 2\pi\delta$, the substitution into the integrals of (52) leads to explicit expressions of the desired solutions $U(n\kappa;z)$ for points $|z|>2\pi\delta$ outside the non-linear layer at each frequency $n\kappa$, $n=1,2,3$.

6. The system of non-linear Sturm-Liouville boundary value problems

The system of non-linear integral equations (52), as well as the problem (31), (C1) – (C4) reduce to a system of non-linear Sturm-Liouville problems.
Indeed, applying the approach described in Yatsyk (2007), Shestopalov & Yatsyk (2007), Kravchenko & Yatsyk (2007), Angermann & Yatsyk (2008), we write the system (52) for arguments z lying in the non-linear layer, i.e. for $|z| \le 2\pi\delta$, in the form

$$\begin{aligned} U(n\kappa;z) + \frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\left[F_{+,n\kappa}(z) + F_{-,n\kappa}(z)\right] &= (\delta_{n1}+\delta_{n3})\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\left[P_{+,n\kappa}(z) + P_{-,n\kappa}(z)\right] \\ &\quad + U^{\mathrm{inc}}(n\kappa;z), \quad |z| \le 2\pi\delta, \quad n=1,2,3, \end{aligned} \tag{54}$$

where

$$\begin{aligned} F_{+,n\kappa}(z) &:= \int_{-2\pi\delta}^{z} \exp(i\Gamma_{n\kappa}(z-z_0)) \times \\ &\quad \times \left[1-\varepsilon_{n\kappa}(z_0,\alpha(z_0),U(\kappa;z_0),U(2\kappa;z_0),U(3\kappa;z_0))\right]U(n\kappa;z_0)\,dz_0, \\ F_{-,n\kappa}(z) &:= \int_{z}^{2\pi\delta} \exp\left(-i\Gamma_{n\kappa}(z-z_0)\right) \times \\ &\quad \times \left[1-\varepsilon_{n\kappa}(z_0,\alpha(z_0),U(\kappa;z_0),U(2\kappa;z_0),U(3\kappa;z_0))\right]U(n\kappa;z_0)\,dz_0, \\ &\qquad\qquad n=1,2,3, \end{aligned}$$

and

$$\begin{aligned} P_{+,\kappa}(z) &:= \int_{-2\pi\delta}^{z} \exp\left(i\Gamma_{\kappa}(z-z_0)\right)\alpha(z_0)U^2(2\kappa;z_0)U^*(3\kappa;z_0)dz_0, \\ P_{-,\kappa}(z) &:= \int_{z}^{2\pi\delta} \exp\left(-i\Gamma_{\kappa}(z-z_0)\right)\alpha(z_0)U^2(2\kappa;z_0)U^*(3\kappa;z_0)dz_0, \\ P_{+,3\kappa}(z) &:= \int_{-2\pi\delta}^{z} \exp\left(i\Gamma_{3\kappa}(z-z_0)\right)\alpha(z_0)\left\{\frac{1}{3}U^3(\kappa;z_0)+U^2(2\kappa;z_0)U^*(\kappa;z_0)\right\}dz_0, \\ P_{-,3\kappa}(z) &:= \int_{z}^{2\pi\delta} \exp\left(-i\Gamma_{3\kappa}(z-z_0)\right)\alpha(z_0)\left\{\frac{1}{3}U^3(\kappa;z_0)+U^2(2\kappa;z_0)U^*(\kappa;z_0)\right\}dz_0. \end{aligned}$$

The integrands and their partial derivatives w.r.t. z are continuous on the set $-2\pi\delta \le z \le 2\pi\delta$, $-2\pi\delta \le z_0 \le 2\pi\delta$. Therefore we may differentiate w.r.t. the argument z by means of the Leibniz rule. Differentiating (54) twice w.r.t. z, we obtain the following system of integro-differential equations:

$$\begin{aligned} &\frac{d^2}{dz^2}U(n\kappa;z) + \frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\left[F''_{+,n\kappa}(z) + F''_{-,n\kappa}(z)\right] \\ &= (\delta_{n1}+\delta_{n3})\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\left[P''_{+,n\kappa}(z) + P''_{-,n\kappa}(z)\right] - \Gamma_{n\kappa}^2 U^{\mathrm{inc}}(n\kappa;z), \quad |z| \le 2\pi\delta, \quad n=1,2,3. \end{aligned} \tag{55}$$

Because of

$$\begin{aligned} F''_{+,n\kappa}(z) + F''_{-,n\kappa}(z) &= i\Gamma_{n\kappa}\left[F'_{+,n\kappa}(z) - F'_{-,n\kappa}(z)\right], \quad n=1,2,3, \\ P''_{+,n\kappa}(z) + P''_{-,n\kappa}(z) &= i\Gamma_{n\kappa}\left[P'_{+,n\kappa}(z) - P'_{-,n\kappa}(z)\right], \quad n=1,3, \end{aligned}$$

where

$$
\begin{array}{rcl}
F'_{+,n\kappa}(z) &=& i\Gamma_{n\kappa}F_{+,n\kappa}(z) + \left[1-\varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\right]U(n\kappa;z),\\
F'_{-,n\kappa}(z) &=& -i\Gamma_{n\kappa}F_{-,n\kappa}(z) - \left[1-\varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\right]U(n\kappa;z),\\
& & n=1,2,3,\\
P'_{+,\kappa}(z) &=& i\Gamma_{\kappa}P_{+,\kappa}(z) + \alpha(z)U^2(2\kappa;z)U^*(3\kappa;z),\\
P'_{-,\kappa}(z) &=& -i\Gamma_{\kappa}P_{-,\kappa}(z) - \alpha(z)U^2(2\kappa;z)U^*(3\kappa;z),\\
P'_{+,3\kappa}(z) &=& i\Gamma_{3\kappa}P_{+,3\kappa}(z) + \alpha(z)\left\{\frac{1}{3}U^3(\kappa;z) + U^2(2\kappa;z)U^*(\kappa;z)\right\},\\
P'_{-,3\kappa}(z) &=& -i\Gamma_{3\kappa}P_{-,3\kappa}(z) - \alpha(z)\left\{\frac{1}{3}U^3(\kappa;z) + U^2(2\kappa;z)U^*(\kappa;z)\right\},
\end{array}
\tag{56}
$$

we see that

$$
\begin{array}{rcl}
F'_{+,n\kappa}(z) - F'_{-,n\kappa}(z) &=& i\Gamma_{n\kappa}\left[F_{+,n\kappa}(z) + F_{-,n\kappa}(z)\right]\\
& & + 2\left[1-\varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\right]U(n\kappa;z),\\
& & n=1,2,3,\\
P'_{+,\kappa}(z) - P'_{-,\kappa}(z) &=& i\Gamma_{\kappa}\left[P_{+,\kappa}(z) + P_{-,\kappa}(z)\right] + 2\alpha(z)U^2(2\kappa;z)U^*(3\kappa;z),\\
P'_{+,3\kappa}(z) - P'_{-,3\kappa}(z) &=& i\Gamma_{3\kappa}\left[P_{+,3\kappa}(z) + P_{-,3\kappa}(z)\right] + 2\alpha(z)\left\{\frac{1}{3}U^3(\kappa;z) + U^2(2\kappa;z)U^*(\kappa;z)\right\}.
\end{array}
$$

Consequently, the system (55) takes the form

$$
\left\{
\begin{array}{rl}
& \frac{d^2}{dz^2}U(n\kappa;z) - \Gamma_{n\kappa}\frac{i(n\kappa)^2}{2}\left[F_{+,n\kappa}(z) + F_{-,n\kappa}(z)\right]\\
& -(n\kappa)^2\left[1-\varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\right]U(n\kappa;z)\\
= & -(\delta_{n1}+\delta_{n3})\frac{i(n\kappa)^2}{2}\Gamma_{n\kappa}\left[P_{+,n\kappa}(z) + P_{-,n\kappa}(z)\right]\\
& -(n\kappa)^2\alpha(z)\left(\delta_{n1}U^2(2\kappa;z)U^*(3\kappa;z) + \delta_{n3}\left\{\frac{1}{3}U^3(\kappa;z) + U^2(2\kappa;z)U^*(\kappa;z)\right\}\right)\\
& -\Gamma_{n\kappa}^2U^{\mathrm{inc}}(n\kappa;z), \quad |z|\le 2\pi\delta, \quad n=1,2,3.
\end{array}
\right.
$$

Making use of the integral representations of the desired solution $\{U(n\kappa;z)\}_{n=1,2,3}$ given by (54), the elimination of the integral terms results in the following system of non-linear second-order ordinary differential equations of Sturm-Liouville type:

$$
\begin{array}{rl}
& \frac{d^2}{dz^2}U(n\kappa;z) + \left\{\Gamma_{n\kappa}^2 - (n\kappa)^2\left[1-\varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\right]\right\}U(n\kappa;z)\\
= & -(n\kappa)^2\alpha(z)\left(\delta_{n1}U^2(2\kappa;z)U^*(3\kappa;z) + \delta_{n3}\left\{\frac{1}{3}U^3(\kappa;z) + U^2(2\kappa;z)U^*(\kappa;z)\right\}\right),\\
& |z|\le 2\pi\delta, \quad n=1,2,3.
\end{array}
\tag{57}
$$

The boundary conditions at $z=\pm 2\pi\delta$ for each of the equations from system (57) are derived from those first-order integro-differential equations, which are obtained by differentiating the integral equations (54) w.r.t. the argument z, i.e.

$$
\begin{array}{rl}
& \frac{d}{dz}U(n\kappa;z) + \frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\left[F'_{+,n\kappa}(z) + F'_{-,n\kappa}(z)\right]\\
= & (\delta_{n1}+\delta_{n3})\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\left[P'_{+,n\kappa}(z) + P'_{-,n\kappa}(z)\right] - i\Gamma_{n\kappa}U^{\mathrm{inc}}(n\kappa;z), \quad |z|\le 2\pi\delta, \quad n=1,2,3.
\end{array}
$$

Because of

$$\begin{aligned} F'_{+,n\kappa}(z)+F'_{-,n\kappa}(z) &= i\Gamma_{n\kappa}\left[F_{+,n\kappa}(z)-F_{-,n\kappa}(z)\right], \quad n=1,2,3,\\ P'_{+,n\kappa}(z)+P'_{-,n\kappa}(z) &= i\Gamma_{n\kappa}\left[P_{+,n\kappa}(z)-P_{-,n\kappa}(z)\right], \quad n=1,3, \end{aligned}$$

(cf. (56)) we get

$$\begin{aligned} &\frac{d}{dz}U(n\kappa;z)+\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}i\Gamma_{n\kappa}\left[F_{+,n\kappa}(z)-F_{-,n\kappa}(z)\right]\\ =\;&(\delta_{n1}+\delta_{n3})\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}i\Gamma_{n\kappa}\left[P_{+,n\kappa}(z)-P_{-,n\kappa}(z)\right]-i\Gamma_{n\kappa}U^{\rm inc}(n\kappa;z), \quad |z|\le 2\pi\delta, \quad n=1,2,3. \end{aligned} \tag{58}$$

Accordingly, at the boundary $z=\pm 2\pi\delta$ the system of integro-differential and integral equations (58), (54) can be represented as

$$\begin{aligned} &\frac{d}{dz}U\left(n\kappa;\begin{Bmatrix}2\pi\delta\\-2\pi\delta\end{Bmatrix}\right)+\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}i\Gamma_{n\kappa}\left[\begin{Bmatrix}F_{+,n\kappa}(2\pi\delta)\\0\end{Bmatrix}-\begin{Bmatrix}0\\F_{-,n\kappa}(-2\pi\delta)\end{Bmatrix}\right]\\ =\;&(\delta_{n1}+\delta_{n3})\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}i\Gamma_{n\kappa}\left[\begin{Bmatrix}P_{+,n\kappa}(2\pi\delta)\\0\end{Bmatrix}-\begin{Bmatrix}0\\P_{-,n\kappa}(-2\pi\delta)\end{Bmatrix}\right]-i\Gamma_{n\kappa}U^{\rm inc}\left(n\kappa;\begin{Bmatrix}2\pi\delta\\-2\pi\delta\end{Bmatrix}\right),\\ &\qquad\qquad n=1,2,3, \end{aligned}$$

and

$$\begin{aligned} &U\left(n\kappa;\begin{Bmatrix}2\pi\delta\\-2\pi\delta\end{Bmatrix}\right)+\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\left[\begin{Bmatrix}F_{+,n\kappa}(2\pi\delta)\\0\end{Bmatrix}-\begin{Bmatrix}0\\F_{-,n\kappa}(-2\pi\delta)\end{Bmatrix}\right]\\ =\;&(\delta_{n1}+\delta_{n3})\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\left[\begin{Bmatrix}P_{+,n\kappa}(2\pi\delta)\\0\end{Bmatrix}-\begin{Bmatrix}0\\P_{-,n\kappa}(-2\pi\delta)\end{Bmatrix}\right]+U^{\rm inc}\left(n\kappa;\begin{Bmatrix}2\pi\delta\\-2\pi\delta\end{Bmatrix}\right),\\ &\qquad\qquad n=1,2,3. \end{aligned}$$

Eliminating from both equations the terms containing the integrals, we obtain the boundary conditions of third kind:

$$\begin{aligned} i\Gamma_{n\kappa}U\left(n\kappa;2\pi\delta\right)-\frac{d}{dz}U\left(n\kappa;2\pi\delta\right) &= 2i\Gamma_{n\kappa}U^{\rm inc}(n\kappa;2\pi\delta),\\ i\Gamma_{n\kappa}U\left(n\kappa;-2\pi\delta\right)+\frac{d}{dz}U\left(n\kappa;-2\pi\delta\right) &= 0, \qquad n=1,2,3. \end{aligned} \tag{59}$$

Therefore, the system of non-linear integral equations (54) (or (52)) according to (57) and (59) is reduced to an equivalent system of non-linear Sturm-Liouville boundary value problems:

$$\begin{aligned} &\frac{d^2}{dz^2}U(n\kappa;z)+\left\{\Gamma_{n\kappa}^2-(n\kappa)^2\left[1-\varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\right]\right\}U(n\kappa;z)\\ =\;&-(n\kappa)^2\alpha(z)\left(\delta_{n1}U^2(2\kappa;z)U^*(3\kappa;z)+\delta_{n3}\left\{\frac{1}{3}U^3(\kappa;z)+U^2(2\kappa;z)U^*(\kappa;z)\right\}\right),\\ &\qquad\qquad |z|\le 2\pi\delta, \end{aligned} \tag{60}$$

$$\begin{aligned} i\Gamma_{n\kappa}U\left(n\kappa;-2\pi\delta\right)+\frac{d}{dz}U\left(n\kappa;-2\pi\delta\right) &= 0,\\ i\Gamma_{n\kappa}U\left(n\kappa;2\pi\delta\right)-\frac{d}{dz}U\left(n\kappa;2\pi\delta\right) &= 2i\Gamma_{n\kappa}U^{\rm inc}(n\kappa;2\pi\delta),\\ &\qquad\qquad n=1,2,3. \end{aligned}$$

We recall that the boundary problem (60) on the interval $|z|\le 2\pi\delta$ can also be obtained by starting from the original problem (31), (C1) – (C4) and the representation of the

desired diffraction field (32), as shown at the end of Section 4. The system of non-linear ordinary differential equations of Sturm-Liouville type follows directly from substituting the representations (32) for the desired solutions, i.e. $\{E_1(n\kappa;y,z) = U(n\kappa;z)\exp(i\phi_{n\kappa}y)\}_{n=1,2,3}$ for $|z| \leq 2\pi\delta$, into the system of equations (31), using the relations $\Gamma_{n\kappa}^2 = (n\kappa)^2 - \phi_{n\kappa}^2$, $n = 1,2,3$, for the longitudinal and transverse propagation constants. The boundary conditions follow from the continuity condition (C3) of the tangential components of the full field of diffraction $\{\mathbf{E}_{\mathrm{tg}}(n\kappa;y,z)\}_{n=1,3}$ $\{\mathbf{H}_{\mathrm{tg}}(n\kappa;y,z)\}_{n=1,3}$ at the boundary $z = \pm 2\pi\delta$ of the non-linear layer:

$$\begin{array}{rclcrcl} U(n\kappa;2\pi\delta) & = & a_{n\kappa}^{\mathrm{scat}} + a_{n\kappa}^{\mathrm{inc}}, & & \dfrac{d}{dz}U(n\kappa;2\pi\delta) & = & i\Gamma_{n\kappa}\left(a_{n\kappa}^{\mathrm{scat}} - a_{n\kappa}^{\mathrm{inc}}\right), \\ U(n\kappa;-2\pi\delta) & = & b_{n\kappa}^{\mathrm{scat}}, & & \dfrac{d}{dz}U(n\kappa;-2\pi\delta) & = & -i\Gamma_{n\kappa}b_{n\kappa}^{\mathrm{scat}}, \quad n = 1,2,3. \end{array} \tag{61}$$

Eliminating in (61) the unknown values of the complex amplitudes $\{a_{n\kappa}^{\mathrm{scat}}\}_{n=1,2,3}$, $\{b_{n\kappa}^{\mathrm{scat}}\}_{n=1,2,3}$ of the scattered field at the boundary $z = \pm 2\pi\delta$ and taking into consideration that $a_{n\kappa}^{\mathrm{inc}} = U^{\mathrm{inc}}(n\kappa;2\pi\delta)$, we arrive at the same boundary conditions as in problem (60).
Thus we have established the equivalence of the non-linear problem (31), (C1) – (C4), of the system of non-linear integral equations (52) and of the system of non-linear boundary-value problems of Sturm-Liouville type (60) (cf. Angermann & Yatsyk (2010), Shestopalov & Yatsyk (2007)).

7. Numerical solution of the non-linear boundary value problem by the finite element method

Using the results given in Angermann & Yatsyk (2008), Angermann & Yatsyk (2010), we can apply the finite element method (FEM) to obtain an approximate solution of the non-linear boundary value problem (60). Let

$$\mathbf{U}(z) := \begin{pmatrix} U(\kappa;z) \\ U(2\kappa;z) \\ U(3\kappa;z) \end{pmatrix},$$

$$\mathbf{F}(z,\mathbf{U}) := \begin{pmatrix} \left\{\Gamma_\kappa^2 - \kappa^2\left[1 - \varepsilon_\kappa(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\right]\right\}U(\kappa;z) \\ \qquad + \ \alpha(z)\kappa^2 U^2(2\kappa;z)U^*(3\kappa;z) \\ \left\{\Gamma_{2\kappa}^2 - (2\kappa)^2\left[1 - \varepsilon_{2\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\right]\right\}U(2\kappa;z) \\ \left\{\Gamma_{3\kappa}^2 - (3\kappa)^2\left[1 - \varepsilon_{3\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\right]\right\}U(3\kappa;z) \\ \qquad + \ \alpha(z)(3\kappa)^2\left\{\frac{1}{3}U^3(\kappa;z) + U^2(2\kappa;z)U^*(\kappa;z)\right\} \end{pmatrix}.$$

Then the system of differential equations in (60) takes the form

$$-\mathbf{U}''(z) = \mathbf{F}(z,\mathbf{U}(z)), \quad z \in \mathcal{I} := (-2\pi\delta, 2\pi\delta). \tag{62}$$

The boundary conditions in (60) can be written as

$$\begin{array}{rcl} \mathbf{U}'(-2\pi\delta) & = & -i\mathbf{G}\mathbf{U}(-2\pi\delta), \\ \mathbf{U}'(2\pi\delta) & = & i\mathbf{G}\mathbf{U}(2\pi\delta) - 2i\mathbf{G}\mathbf{a}^{\mathrm{inc}}, \end{array} \tag{63}$$

where

$$\mathbf{G} := \begin{pmatrix} \Gamma_\kappa & 0 & 0 \\ 0 & \Gamma_{2\kappa} & 0 \\ 0 & 0 & \Gamma_{3\kappa} \end{pmatrix} \qquad \text{and} \qquad \mathbf{a}^{\mathrm{inc}} := \begin{pmatrix} a_\kappa^{\mathrm{inc}} \\ a_{2\kappa}^{\mathrm{inc}} \\ a_{3\kappa}^{\mathrm{inc}} \end{pmatrix}.$$

Taking an arbitrary complex-valued vector function $\mathbf{v}:\ [-2\pi\delta,2\pi\delta]\to\mathbb{C}^3$, $\mathbf{v}=\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}$, multiplying the vector differential equation (62) by the complex conjugate $\mathbf{v}^*$ and integrating w.r.t. z over the interval $\mathcal{I}$, we arrive at the equation

$$-\int_{\mathcal{I}}\mathbf{U}''\cdot\mathbf{v}^*\,dz=\int_{\mathcal{I}}\mathbf{F}(z,\mathbf{U})\cdot\mathbf{v}^*\,dz.$$

Integrating the left-hand side of this equation by parts and using the boundary conditions (63), we obtain:

$$\begin{array}{rcl}-\int_{\mathcal{I}}\mathbf{U}''\cdot\mathbf{v}^*\,dz & = & \int_{\mathcal{I}}\mathbf{U}'\cdot\mathbf{v}^*\,dz-(\mathbf{U}'\cdot\mathbf{v}^*)(2\pi\delta)+(\mathbf{U}'\cdot\mathbf{v}^*)(-2\pi\delta)\\ & = & \int_{\mathcal{I}}\mathbf{U}'\cdot\mathbf{v}^{*\prime}dz-i\left[((\mathbf{GU})\cdot\mathbf{v}^*)(2\pi\delta)+((\mathbf{GU})\cdot\mathbf{v}^*)(-2\pi\delta)\right]\\ & & +2i(\mathbf{Ga}^{\mathrm{inc}})\cdot\mathbf{v}^*(2\pi\delta).\end{array}$$

Now we consider the complex Sobolev space $H^1(\mathcal{I})$ consisting of functions with values in $\mathbb{C}$, which, together with their weak derivatives belong to $L_2(\mathcal{I})$. For $\mathbf{w},\mathbf{v}\in\left[H^1(\mathcal{I})\right]^3$, we introduce the following forms:

$$\begin{array}{rcl}a(\mathbf{w},\mathbf{v}) & := & \int_{\mathcal{I}}\mathbf{w}'\cdot\mathbf{v}^{*\prime}dz-i\left[((\mathbf{Gw})\cdot\mathbf{v}^*)(2\pi\delta)+((\mathbf{Gw})\cdot\mathbf{v}^*)(-2\pi\delta)\right],\\ b(\mathbf{w},\mathbf{v}) & := & \int_{\mathcal{I}}\mathbf{F}(z,\mathbf{w})\cdot\mathbf{v}^*dz-2i(\mathbf{Ga}^{\mathrm{inc}})\cdot\mathbf{v}^*(2\pi\delta).\end{array}$$

So we arrive at the following weak formulation of boundary value problem (60):

Find $\mathbf{U}\in\left[H^1(\mathcal{I})\right]^3$ such that

$$a(\mathbf{U},\mathbf{v})=b(\mathbf{U},\mathbf{v})\quad\forall\mathbf{v}\in\left[H^1(\mathcal{I})\right]^3.\tag{64}$$

Based on the variational equation (64), we obtain the numerical method. We consider N nodes $\{z_i\}_{i=1}^N$ such that $-2\pi\delta=:z_1<z_2<\ldots<z_{N-1}<z_N:=2\pi\delta$, and define the intervals $\mathcal{I}_i:=(z_i,z_{i+1})$ with the lengths $h_i:=z_{i+1}-z_i$ and the parameter $h:=\max_{i\in\{1,\ldots,N-1\}}h_i$. Then, for $i\in\{1,\ldots,N\}$ we introduce the basis functions $\psi_i:\ [-2\pi\delta,2\pi\delta]\to\mathbb{R}$ by the formula

$$\psi_i(z):=\begin{cases}(z-z_{i-1})/h_{i-1}, & z\in\mathcal{I}_{i-1}\text{ and }i\geq 2,\\ (z_{i+1}-z)/h_i, & z\in\mathcal{I}_i\text{ and }i\leq N-1,\\ 0, & \text{otherwise}\end{cases}$$

and the corresponding space $V_h:=\left\{v_h=\sum_{i=1}^N\lambda_i\psi_i:\ \lambda_i\in\mathbb{C}\right\}$ (defined by a set of all linear combinations of the basis functions). It is well-known that $V_h\subset H^1(\mathcal{I})$ (cf. Samarskij & Gulin (2003)). Therefore the following discrete finite element formulation of the problem (64) is well-defined (see Angermann & Yatsyk (2008), Samarskij & Gulin (2003)):

Find $\mathbf{U}_h\subset V_h^3$ such that

$$a(\mathbf{U}_h,\mathbf{v}_h)=b_h(\mathbf{U}_h,\mathbf{v}_h)\quad\forall\mathbf{v}_h:=\begin{pmatrix}v_{h1}\\v_{h2}\\v_{h3}\end{pmatrix}\in V_h^3.\tag{65}$$

The non-linear discrete form b_h is a slight modification of the right-hand side b of the problem (64) defined as follows:

$$b_h(\mathbf{w}_h,\mathbf{v}_h):=\int_{\mathcal{I}}\left[\mathbf{F}_h^{(L)}(z,\mathbf{w}_h)+\mathbf{F}_h^{(NL)}(z,\mathbf{w}_h)\right]\cdot\mathbf{v}_h^*dz-2i(\mathbf{Ga}^{\mathrm{inc}})\cdot\mathbf{v}_h^*(2\pi\delta),$$

where

$$\mathbf{F}_h^{(L)}(z,\mathbf{w}_h) \quad := \quad \begin{pmatrix} \{\Gamma_\kappa^2 - \kappa^2(1-\varepsilon^{(L)})\}\, w_1 \\ \{\Gamma_{2\kappa}^2 - (2\kappa)^2(1-\varepsilon^{(L)})\}\, w_2 \\ \{\Gamma_{3\kappa}^2 - (3\kappa)^2(1-\varepsilon^{(L)})\}\, w_3 \end{pmatrix},$$

$$\mathbf{F}_h^{(NL)}(z,\mathbf{w}_h) \quad := \quad \begin{pmatrix} \kappa^2 \sum_{i=1}^N \left[\varepsilon_\kappa^{(NL)}(z_i,\, \alpha(z_i), w_{1i}, w_{2i}, w_{3i})\, w_{1i} \right. \\ \qquad\qquad \left. +\ \alpha(z_i) w_{2i}^2 w_{3i}^*\right] \psi_i \\ (2\kappa)^2 \sum_{i=1}^N \varepsilon_{2\kappa}^{(NL)}(z_i,\, \alpha(z_i), w_{1i}, w_{2i}, w_{3i})\, w_{2i}\, \psi_i \\ (3\kappa)^2 \sum_{i=1}^N \left[\varepsilon_{3\kappa}^{(NL)}(z_i,\, \alpha(z_i), w_{1i}, w_{2i}, w_{3i})\, w_{3i} \right. \\ \qquad\qquad \left. +\ \alpha(z_i)\left\{\tfrac{1}{3} w_{1i}^3 + w_{2i}^2 w_{1i}^*\right\}\right] \psi_i \end{pmatrix}.$$

In fact, the problem (65) reduces to solving a non-linear system of algebraic equations w.r.t. $3N$ complex scalars.

As in Angermann & Yatsyk (2008) the weak formulation (64) and the discrete formulation (65) can be used to prove, under certain assumptions, the existence and uniqueness of the solutions $\mathbf{U} \in \left[H^1(\mathcal{I})\right]^3$ and $\mathbf{U}_h \in V_h^3$, respectively. Furthermore, the convergence of the finite element solution to the weak solution can be established.

8. Third harmonic generation and resonant scattering of a strong electromagnetic field by the non-linear structure. A numerical algorithm for solving systems of non-linear integral equations

Consider the excitation of the non-linear structure by a strong electromagnetic field at the basic frequency κ only (see (30)), i.e.

$$\{E_1^{\rm inc}(\kappa;q) \neq 0, \quad E_1^{\rm inc}(2\kappa;q) = 0, \quad E_1^{\rm inc}(3\kappa;q) = 0\}, \quad \text{where} \quad \{a_\kappa^{\rm inc} \neq 0, \quad a_{2\kappa}^{\rm inc} = a_{3\kappa}^{\rm inc} = 0\}.$$

In this case, the number of equations in the system of non-linear boundary-value problems (31), (C1) – (C4) and in the equivalent system of Sturm-Liouville problems (60), and the number of non-linear integral equations in the system (52) can be reduced (cf. Angermann & Yatsyk (2010)). As noted above, the second equation in each of the systems (31), (60) and (52), corresponding to a problem at the double frequency 2κ with a trivial right-hand side, can be eliminated by setting $E_1(\mathbf{r},2\kappa) := 0$. The dielectric permittivity of the non-linear layer depends on the component $U(\kappa;z)$ of the scattered field and on the component $U(3\kappa;z)$ of the generated field, i.e. the expression (29) simplifies to

$$\begin{aligned} & \varepsilon_{n\kappa}(z,\alpha(z),E_1(\mathbf{r},\kappa),0,E_1(\mathbf{r},3\kappa)) = \varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(3\kappa;z)) \\ =: \ & \varepsilon^{(L)}(z) + \varepsilon_{n\kappa}^{(NL}(\alpha(z),U(\kappa;z),U(3\kappa;z)) \\ = \ & \varepsilon^{(L)}(z) + \alpha(z)\left[|U(\kappa;z)|^2 + |U(3\kappa;z)|^2\right] \\ & + \delta_{n,1}\alpha(z)|U(\kappa;z)||U(3\kappa;z)|\exp\left[i\left\{-3\arg U(\kappa;z) + \arg U(3\kappa;z)\right\}\right], \quad n = 1,3. \end{aligned} \tag{66}$$

Now we discuss the numerical realisation of the approach based on the non-linear integral equations (52). In the case under consideration, the problem is reduced to finding solutions to one-dimensional non-linear integral equations (along the height $z \in [-2\pi\delta, 2\pi\delta]$ of the structure) w.r.t. the components $U(n\kappa;z)$, $U(3n\kappa;z)$:

$$\begin{cases} U(\kappa;z)+\dfrac{i\kappa^2}{2\Gamma_\kappa}\displaystyle\int_{-2\pi\delta}^{2\pi\delta}\exp(i\Gamma_\kappa|z-z_0|)\left[1-\varepsilon_\kappa\left(z_0,\alpha(z_0),U(\kappa;z_0),U(3\kappa;z_0)\right)\right]U(\kappa;z_0)dz_0 \\ \quad =U^{\mathrm{inc}}(\kappa;z), & |z|\le 2\pi\delta, \\ U(3\kappa;z)+\dfrac{i(3\kappa)^2}{2\Gamma_{3\kappa}}\displaystyle\int_{-2\pi\delta}^{2\pi\delta}\exp(i\Gamma_{3\kappa}|z-z_0|)\left[1-\varepsilon_{3\kappa}(z_0,\alpha(z_0),U(\kappa;z_0),U(3\kappa;z_0))\right]U(3\kappa;z_0)dz_0 \\ \quad =\dfrac{i(3\kappa)^2}{6\Gamma_{3\kappa}}\displaystyle\int_{-2\pi\delta}^{2\pi\delta}\exp(i\Gamma_{3\kappa}|z-z_0|)\alpha(z_0)U^3(\kappa;z_0)dz_0, & |z|\le 2\pi\delta, \end{cases} \tag{67}$$

where $U^{\mathrm{inc}}(\kappa;z)=a_\kappa^{\mathrm{inc}}\exp\left[-i\Gamma_\kappa(z-2\pi\delta)\right]$.
The desired solution of the diffraction problem (31), (C1) – (C4) can be represented as follows (cf. (32)):

$$\begin{aligned} &E_1(n\kappa;y,z)=U(n\kappa;z)\exp(i\phi_{n\kappa}y) \\ &=\begin{cases}\delta_{n1}a_{n\kappa}^{\mathrm{inc}}\exp(i(\phi_{n\kappa}y-\Gamma_{n\kappa}(z-2\pi\delta)))+a_{n\kappa}^{\mathrm{scat}}\exp(i(\phi_{n\kappa}y+\Gamma_{n\kappa}(z-2\pi\delta))), & z>2\pi\delta,\\ U(n\kappa;z)\exp(i\phi_{n\kappa}y), & |z|\le 2\pi\delta,\\ b_{n\kappa}^{\mathrm{scat}}\exp(i(\phi_{n\kappa}y-\Gamma_{n\kappa}(z+2\pi\delta))), & z<-2\pi\delta,\end{cases} \\ &\quad n=1,3, \end{aligned} \tag{68}$$

where $U(\kappa;z)$, $U(3\kappa;z)$, $|z|\le 2\pi\delta$, are the solutions of the system (67). According to (53) we determine the values of complex amplitudes $\left\{a_{n\kappa}^{\mathrm{scat}},b_{n\kappa}^{\mathrm{scat}}:\ n=1,3\right\}$ in (68) for the scattered and generated fields by means of the formulas

$$U(n\kappa;2\pi\delta)=\delta_{n1}a_{n\kappa}^{\mathrm{inc}}+a_{n\kappa}^{\mathrm{scat}},\quad U(n\kappa;-2\pi\delta)=b_{n\kappa}^{\mathrm{scat}},\quad n=1,3. \tag{69}$$

The solution of the system of non-linear integral equations (67) can be approximated numerically by the help of an iterative method. The proposed algorithm is based on the application of a quadrature rule to each of the non-linear integral equations of the system (67). The resulting system of complex non-linear inhomogeneous algebraic equations is solved by a block-iterative method, cf. Yatsyk (September 21-24, 2009), Yatsyk (June 21-26, 2010).
Thus, using Simpson's quadrature rule, the system of non-linear integral equations (67) reduces to a system of non-linear algebraic equations of the second kind:

$$\begin{cases} \left(\mathbf{I}-\mathbf{B}_\kappa(\mathbf{U}_\kappa,\mathbf{U}_{3\kappa})\right)\mathbf{U}_\kappa &= \mathbf{U}_\kappa^{\mathrm{inc}}, \\ \left(\mathbf{I}-\mathbf{B}_{3\kappa}(\mathbf{U}_\kappa,\mathbf{U}_{3\kappa})\right)\mathbf{U}_{3\kappa} &= \mathbf{C}_{3\kappa}\left(\mathbf{U}_\kappa\right), \end{cases} \tag{70}$$

where, as in Section 7, $\{z_i\}_{i=1}^N$ is a discrete set of nodes $-2\pi\delta=:z_1<z_2<\ldots<z_n<\ldots<z_N=:2\pi\delta$.
$\mathbf{U}_{p\kappa}:=\{U_n(p\kappa)\}_{n=1}^N\approx\{U(p\kappa;z_n)\}_{n=1}^N$ denotes the vector of the unknown approximate solution values corresponding to the frequencies $p\kappa$, $p=1,3$. The matrices are of the form

$$\mathbf{B}_{p\kappa}(\mathbf{U}_\kappa,\mathbf{U}_{3\kappa})=\{A_mK_{nm}(p\kappa,\mathbf{U}_\kappa,\mathbf{U}_{3\kappa})\}_{n,m=1}^N$$

with entries

$$\begin{aligned} K_{nm}(p\kappa,\mathbf{U}_\kappa,\mathbf{U}_{3\kappa}) \quad := \quad & -\frac{i(p\kappa)^2}{2\Gamma_{p\kappa}}\exp\left(i\Gamma_{p\kappa}|z_n-z_m|\right)\Big[1-\Big\{\varepsilon^{(L)}(z_m) \\ & +\alpha(z_m)\left(|U_m(\kappa)|^2+|U_m(3\kappa)|^2\right. \\ & +\ \ \delta_{p1}\,|U_m(\kappa)||U_m(3\kappa)|\exp\left\{i\left[-3\arg U_m(\kappa)+\arg U_m(3\kappa)\right]\right\}\Big)\Big\}\Big]. \end{aligned}$$

The numbers A_m are the coefficients determined by the quadrature rule, $\mathbf{I} := \{\delta_{nm}\}_{n,m=1}^{N}$ is the identity matrix, and δ_{nm} is Kronecker's symbol.
The right-hand side of (70) is defined by

$$\begin{aligned} \mathbf{U}_{\kappa}^{\text{inc}} &:= \left\{a_{\kappa}^{\text{inc}} \exp\left[-i\Gamma_{\kappa}(z_n - 2\pi\delta)\right]\right\}_{n=1}^{N}, \\ \mathbf{C}_{3\kappa}\left(\mathbf{U}_{\kappa}\right) &:= \left\{\frac{i(3\kappa)^2}{6\,\Gamma_{3\kappa}} \sum_{m=1}^{N} A_m \exp\left(i\Gamma_{3\kappa}|z_n - z_m|\right) \alpha\left(z_m\right) U_m^3(\kappa)\right\}_{n=1}^{N}. \end{aligned}$$

Given a relative error tolerance $\xi > 0$, the approximate solution of (70) is obtained by means of the following iterative method:

$$\left\{ \begin{array}{c} \left\{\left[\mathbf{I} - \mathbf{B}_{\kappa}\left(\mathbf{U}_{\kappa}^{(s-1)}, \mathbf{U}_{3\kappa}^{(S_{3q})}\right)\right] \mathbf{U}_{\kappa}^{(s)} = \mathbf{U}_{\kappa}^{\text{inc}}\right\}_{s=1}^{S_q:\, \|\mathbf{U}_{\kappa}^{(S_q)} - \mathbf{U}_{\kappa}^{(S_q-1)}\| / \|\mathbf{U}_{\kappa}^{(S_q)}\| < \xi} \\ \left\{\left[\mathbf{I} - \mathbf{B}_{3\kappa}\left(\mathbf{U}_{\kappa}^{(S_q)}, \mathbf{U}_{3\kappa}^{(s-1)}\right)\right] \mathbf{U}_{3\kappa}^{(s)} = \mathbf{C}_{3\kappa}\left(\mathbf{U}_{\kappa}^{(S_q)}\right)\right\}_{s=1}^{S_{3q}:\, \|\mathbf{U}_{3\kappa}^{(S_{3q})} - \mathbf{U}_{3\kappa}^{(S_{3q}-1)}\| / \|\mathbf{U}_{3\kappa}^{(S_{3q})}\| < \xi} \end{array} \right\}_{q=1}^{Q}, \quad (71)$$

where the terminating index $Q \in \mathbb{N}$ is defined by the requirement

$$\max\left\{\|\mathbf{U}_{\kappa}^{(Q)} - \mathbf{U}_{\kappa}^{(Q-1)}\| / \|\mathbf{U}_{\kappa}^{(Q)}\|, \|\mathbf{U}_{3\kappa}^{(Q)} - \mathbf{U}_{3\kappa}^{(Q-1)}\| / \|\mathbf{U}_{3\kappa}^{(Q)}\|\right\} < \xi.$$

We mention that, as in Yatsyk (2006), Shestopalov & Yatsyk (2007), a sufficient condition for convergence of the iterative process (71) can be derived. Similarly, under appropriate assumptions, a condition for existence and uniqueness of the solution of the problem can be obtained.

9. Numerical analysis. Resonant scattering of waves and the generation of the third harmonic

We consider a non-linear dielectric layered structure (see Fig. 1), the dielectric permittivity

$$\varepsilon_{n\kappa}(z, \alpha(z), U(\kappa; z), U(3\kappa; z)) = \varepsilon^{(L)} + \varepsilon_{n\kappa}^{(NL)}$$

of which is given by (29), where

$$\left\{\varepsilon^{(L)}(z), \alpha(z)\right\} = \left\{ \begin{array}{ll} \left\{\varepsilon^{(L)} = 16,\ \alpha = \alpha_1\right\}, & z \in [-2\pi\delta, z_1 = -2\pi\delta/3) \\ \left\{\varepsilon^{(L)} = 64,\ \alpha = \alpha_2\right\}, & z \in [z_1 = -2\pi\delta/3,\ z_2 = 2\pi\delta/3] \\ \left\{\varepsilon^{(L)} = 16,\ \alpha = \alpha_3\right\}, & z \in (z_2 = 2\pi\delta/3,\ 2\pi\delta] \end{array} \right\},$$

$\alpha_1 = \alpha_3 = 0.01$, $\alpha_2 = -0.01$, $\delta = 0.5$. The excitation frequency is given by $\kappa = 0.25$, and the angle of incidence of the plane wave at the basic frequency κ is $\varphi_{\kappa} \in [0^{\circ}, 90^{\circ})$.
By $W_{n\kappa} = |a_{n\kappa}^{\text{scat}}|^2 + |b_{n\kappa}^{\text{scat}}|^2$ we denote the total energy of the scattered and generated fields at the frequencies $n\kappa$, $n = 1,3$. Thus W_{κ} is the total energy scattered at the frequency κ of excitation, and $W_{3\kappa}$ is the total energy generated at the frequency 3κ. Fig. 2 (left) shows the dependence of $W_{3\kappa}/W_{\kappa}$ on the angle of incidence φ_{κ} and on the amplitude a_{κ}^{inc} of the incident field. It describes the portion of energy generated in the third harmonic by the non-linear layer when a plane wave with angle of incidence φ_{κ} and amplitude a_{κ}^{inc} is passing the layer.

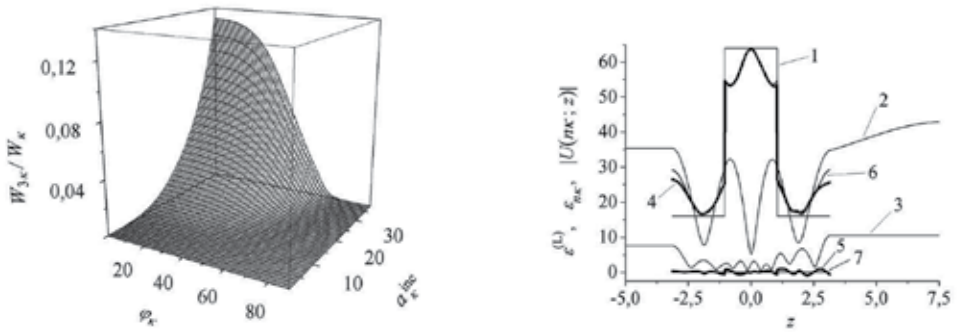


Fig. 2. The portion of energy generated in the third harmonic (left) and some graphs describing the properties of the structure at $a_\kappa^{\mathrm{inc}} = 38$ and $\varphi_\kappa = 0^\circ$ (right): #1 ... $\varepsilon^{(L)}$, #2 ... $|U(\kappa;z)|$, #3 ... $|U(3\kappa;z)|$, #4 ... $\mathfrak{Re}(\varepsilon_\kappa)$, #5 ... $\mathfrak{Im}(\varepsilon_\kappa)$, #6 ... $\mathfrak{Re}(\varepsilon_{3\kappa})$, #7 ... $\mathfrak{Im}(\varepsilon_{3\kappa}) \equiv 0$

In particular, $W_{3\kappa}/W_\kappa = 0.132$ at $a_\kappa^{\mathrm{inc}} = 38$, i.e. $W_{3\kappa}$ amounts to 13.2% of the total energy W_κ scattered at the frequency of excitation κ.

Fig. 2 (right) shows the absolute values of the amplitudes of the full scattered field (total diffraction field) $|U(\kappa;z)|$ at the frequency of excitation κ (graph #2) and of the generated field $|U(3\kappa;z)|$ at the frequency 3κ (graph #3). The values $|U(\kappa;z)|$ and $|U(3\kappa;z)|$ are given in the non-linear layered structure ($|z| \le 2\pi\delta$) and outside it (i.e. in the zones of reflection $z > 2\pi\delta$ and transmission $z < -2\pi\delta$). Fig. 2 (right) also displays some graphs characterising the scattering and generation properties of the non-linear structure. Graph #1 illustrates the value of the linear part $\varepsilon^{(L)}$ of the permittivity of the non-linear layered structure. Graphs #4 and #5 show the real and imaginary part of the permittivity at the frequency of excitation, while graphs #6 and #7 display the corresponding values at the generation frequency.

Figs. 3, 4 and 5 show the numerical results obtained for the scattered and the generated fields and for the non-linear dielectric permittivity in dependence on the amplitude a_κ^{inc} at normal incidence $\varphi_\kappa = 0^\circ$ of the plane wave.

Fig. 3 shows the graphs of $|U_\kappa\left[a_\kappa^{\mathrm{inc}},z\right]|$ and $|U_{3\kappa}\left[a_\kappa^{\mathrm{inc}},z\right]|$ demonstrating the behaviour of the scattered and the generated fields, $|U(\kappa;z)|$ and $|U(3\kappa;z)|$, in the non-linear layered

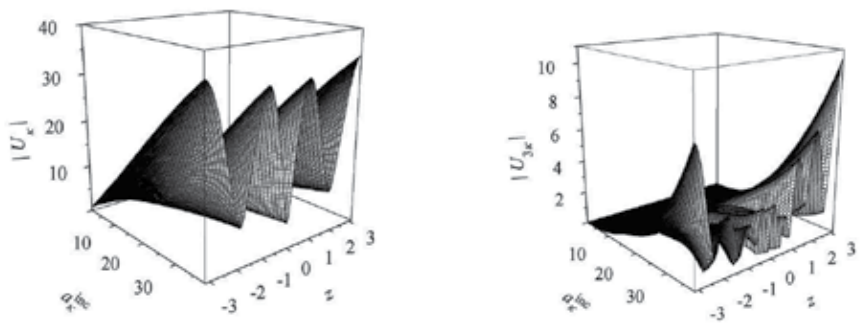


Fig. 3. Graphs of the scattered and generated fields in the non-linear layered structure for $\varphi_\kappa = 0^\circ$: $|U_\kappa\left[a_\kappa^{\mathrm{inc}},z\right]|$ at $\kappa = 0.25$ (left), $|U_{3\kappa}\left[a_\kappa^{\mathrm{inc}},z\right]|$ at $3\kappa = 0.75$ (right)

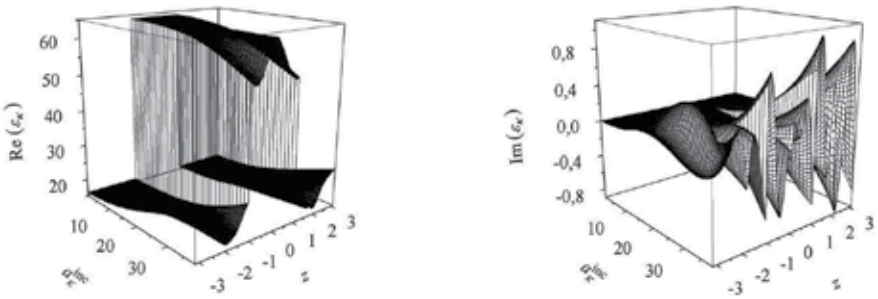


Fig. 4. Graphs of the permittivity at the frequency of excitation $\kappa = 0.25$ at normal incidence of the plane wave $\varphi_\kappa = 0°$: $\mathfrak{Re}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}},z\right]\right)$ (left), $\mathfrak{Im}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}},z\right]\right)$ (right)

structure in dependence on an increasing amplitude a_κ^{inc} at normal incidence $\varphi_\kappa = 0°$ of the plane wave of the frequency $\kappa = 0.25$. According to (66), the non-linear parts $\varepsilon_{n\kappa}^{(NL)}$ of the dielectric permittivity at each frequency κ and 3κ depend on the values $U_\kappa := U(\kappa;z)$ and $U_{3\kappa} := U(3\kappa;z)$ of the fields. The variation of the non-linear parts $\varepsilon_{n\kappa}^{(NL)}$ of the dielectric permittivity for an increasing amplitude a_κ^{inc} of the incident field are illustrated by the behaviour of $\mathfrak{Re}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}},z\right]\right)$ (Fig. 4 (left)) and $\mathfrak{Im}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}},z\right]\right)$ (Fig. 4 (right)) at the frequency κ, and by $\varepsilon_{3\kappa}\left[a_\kappa^{\mathrm{inc}},z\right]$ at the triple frequency 3κ (Fig. 5 (left)).
In Fig. 4 (right) the graph of $\mathfrak{Im}(\varepsilon_\kappa)$ for a given amplitude a_κ^{inc} (denoted by $\mathfrak{Im}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}},z\right]\right)$) characterises the loss of energy in the non-linear medium (at the frequency of excitation κ) caused by the generation of the electromagnetic field of the third harmonic (at the frequency 3κ). In our case $\mathfrak{Im}\left[\varepsilon^{(L)}(z)\right] = 0$ and $\mathfrak{Im}\left[\alpha(z)\right] = 0$, therefore, according to (66),

$$\mathfrak{Im}(\varepsilon_\kappa) = \alpha(z)|U(\kappa;z)||U(3\kappa;z)|\mathfrak{Im}\left(\exp\left[i\left\{-3\arg U(\kappa;z) + \arg U(3\kappa;z)\right\}\right]\right). \tag{72}$$

Fig. 4 (right) shows that the third harmonic generation is insignificant, i.e. $U(3\kappa;z) \approx 0$, if the non-linear structure is excited by a weak field (cf. also Figs. 4 (left), 5 and 3). In this case, for a small value of $|a_\kappa^{\mathrm{inc}}|$ in Fig. 4 (right) we observe a small amplitude of the function $\mathfrak{Im}(\varepsilon_\kappa)$, i.e. $|\mathfrak{Im}(\varepsilon_\kappa)| \approx 0$. The increase of $|a_\kappa^{\mathrm{inc}}|$ corresponds to a strong field excitation and leads to the generation of a third harmonic field $U(3\kappa;z)$. In this case, the variation of the absolute

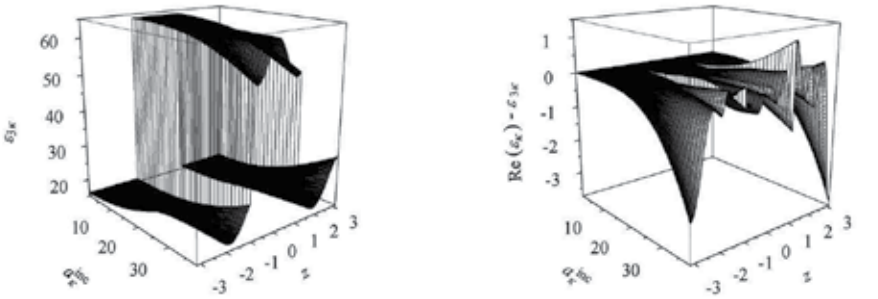


Fig. 5. Graph of the dielectric permittivity $\varepsilon_{3\kappa}\left[a_\kappa^{\mathrm{inc}},z\right]$ at the triple frequency $3\kappa = 0.75$ for $\varphi_\kappa = 0°$ (left), behaviour of $\mathfrak{Re}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}},z\right]\right) - \varepsilon_{3\kappa}\left[a_\kappa^{\mathrm{inc}},z\right]$ (right)

values $|U(\kappa;z)|$, $|U(3\kappa;z)|$ of the scattered and generated fields increase, see Fig. 3. Fig. 4 (right) shows that the values of $\mathfrak{Im}(\varepsilon_\kappa)$ may be positive or negative along the height of the non-linear layer, i.e. in the interval $z \in [-2\pi\delta, 2\pi\delta]$. The zero values of $\mathfrak{Im}(\varepsilon_\kappa)$ are determined by the phase relation between the scattered and the generated fields $U(\kappa;z)$, $U(3\kappa;z)$ in the non-linear layer, see (72),

$$-3\arg U(\kappa;z) + \arg U(3\kappa;z) = p\pi, \quad p = 0, \pm 1, \ldots$$

We mention that the behaviour of both the quantities $\mathfrak{Im}(\varepsilon_\kappa)$ and

$$\mathfrak{Re}(\varepsilon_\kappa) - \varepsilon_{3\kappa} = \alpha(z)|U(\kappa;z)||U(3\kappa;z)|\mathfrak{Re}\left(\exp\left[i\left\{-3\arg U(\kappa;z) + \arg U(3\kappa;z)\right\}\right]\right)$$

plays a role in the process of third harmonic generation because of the presence of the last term in (66). Fig. 5 (right) shows the graph describing the behaviour of $\mathfrak{Re}\left(\varepsilon_\kappa\left[a_\kappa^{\text{inc}}, z\right]\right) - \varepsilon_{3\kappa}\left[a_\kappa^{\text{inc}}, z\right]$.

In order to describe the scattering and generation properties of the non-linear structure in the zones of reflection $z > 2\pi\delta$ and transmission $z < -2\pi\delta$, we introduce the following notation:

$$R_{n\kappa} := |a_{n\kappa}^{\text{scat}}|^2 / |a_\kappa^{\text{inc}}|^2 \quad \text{and} \quad T_{n\kappa} := |b_{n\kappa}^{\text{scat}}|^2 / |a_\kappa^{\text{inc}}|^2.$$

The quantities $R_{n\kappa}$, $T_{n\kappa}$ represent the portions of energy of the reflected and the transmitted waves (at the excitation frequency κ), or the portions of energy of the generated waves in the zones of reflection and transmission (at the frequency 3κ), with respect to the energy of the incident field (at the frequency κ). We call them *reflection, transmission* or *generation coefficients* of the waves w.r.t. the intensity of the excitation field.

We note that in the considered case of the excitation $\{a_\kappa^{\text{inc}} \neq 0, a_{2\kappa}^{\text{inc}} = 0, a_{3\kappa}^{\text{inc}} = 0\}$ and for non-absorbing media with $\mathfrak{Im}\left[\varepsilon^{(L)}(z)\right] = 0$, the energy balance equation

$$R_\kappa + T_\kappa + R_{3\kappa} + T_{3\kappa} = 1$$

is satisfied. This equation represents the law of conservation of energy (Shestopalov & Sirenko (1989), Vainstein (1988)). It can be obtained by writing the energy conservation law for each frequency κ and 3κ, adding the resulting equations and taking into consideration the fact that the loss of energy at the frequency κ (spent for the generation of the third harmonic) is equal to the amount of energy generated at the frequency 3κ.

The scattering and generation properties of the non-linear structure are presented in Figs. 6 – 8. We consider the following range of parameters of the excitation field: the angle $\varphi_\kappa \in [0^\circ, 90^\circ)$, the amplitude of the incident plane wave $a_\kappa^{\text{inc}} \in [1, 38]$ at the frequency $\kappa = 0.25$. The graphs show the dynamics of the scattering ($R_\kappa\left[\varphi_\kappa, a_\kappa^{\text{inc}}\right]$, $T_\kappa\left[\varphi_\kappa, a_\kappa^{\text{inc}}\right]$, see Fig. 6) and generation ($R_{3\kappa}\left[\varphi_\kappa, a_\kappa^{\text{inc}}\right]$, $T_{3\kappa}\left[\varphi_\kappa, a_\kappa^{\text{inc}}\right]$, see Fig. 7) properties of the structure.

Fig. 8 shows cross sections of the graphs depicted in Figs. 6 – 7 by the planes $\varphi_\kappa = 0^\circ$ and $a_\kappa^{\text{inc}} = 38$. We see that increasing the amplitude of the excitation field of the non-linear layer leads to the third harmonic generation (Fig. 8 (left)). In the range $29 < a_\kappa^{\text{inc}} \leq 38$ (i.e. right from the intersection of the graphs #1 and #3 in Fig. 8 (left)) we see that $R_{3\kappa} > R_\kappa$. In this case, $0.053 < W_{3\kappa}/W_\kappa \leq 0.132$, cf. Fig. 2. If $34 < a_\kappa^{\text{inc}} \leq 38$ (i.e. right from the intersection of the graphs #1 and #4 in Fig. 8 (left)) the field generated at the triple frequency in the zones of reflection and transmission is stronger than the reflected field at the excitation frequency κ: $R_{3\kappa} > T_{3\kappa} > R_\kappa$. Here, $0.088 < W_{3\kappa}/W_\kappa \leq 0.132$, cf. Fig. 2.

Fig. 8 (right) shows the dependence of the coefficients of the scattered and generated waves on the angle of incidence $\varphi_\kappa \in [0^\circ, 90^\circ)$ of a plane wave with a constant amplitude $a_\kappa^{\text{inc}} = 38$

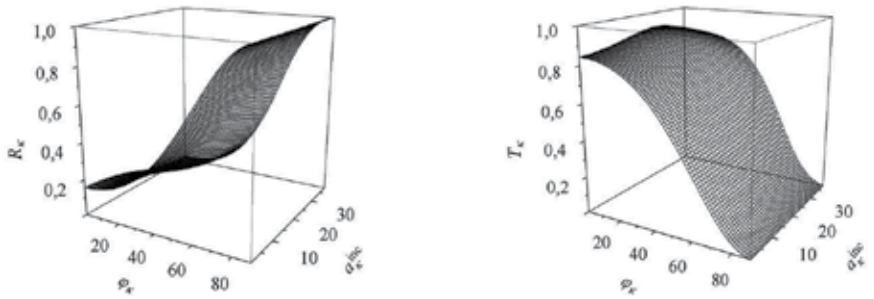


Fig. 6. The scattering properties of the non-linear structure at the excitation frequency $\kappa = 0.25$: $R_\kappa\left[\varphi_\kappa, a_\kappa^{\text{inc}}\right]$ (left), $T_\kappa\left[\varphi_\kappa, a_\kappa^{\text{inc}}\right]$ (right)

of the incident field. It is seen that an increasing angle φ_κ leads to a weakening of the third harmonic generation. In the range of angles $0^\circ \le \varphi_\kappa < 21^\circ$ (i.e. left from the intersection of the graphs #1 and #4 in Fig. 8 (right)) we see that $T_{3\kappa} > R_\kappa$. In this case, $0.125 < W_{3\kappa}/W_\kappa \le 0.132$, cf. Fig. 2. The value of the coefficient of the third harmonic generation in the zone of reflection exceeds the value of the reflection coefficient at the excitation frequency, i.e. $R_{3\kappa} > R_\kappa$, in the range of angles $0^\circ \le \varphi_\kappa < 27^\circ$ (i.e. left from the intersection of the graphs #1 and #3 in Fig. 8 (right)). Here, according to Fig. 2, $0.117 < W_{3\kappa}/W_\kappa \le 0.132$. We mention that, at the normal incidence $\varphi_\kappa = 0^\circ$ of a plane wave with amplitude $a_\kappa^{\text{inc}} = 38$, the coefficients of generation in the zones of reflection $R_{3\kappa}\left[\varphi_\kappa = 0^\circ, a_\kappa^{\text{inc}} = 38\right] = 0.076$ and transmission $T_{3\kappa}\left[\varphi_\kappa = 0^\circ, a_\kappa^{\text{inc}} = 38\right] = 0.040$ reach their maximum values, see Fig.s 7 and 8. In this case, the coefficients describing the portion of reflected and transmitted waves at the frequency of excitation $\kappa = 0.25$ of the structure take the following values: $R_\kappa\left[\varphi_\kappa = 0^\circ, a_\kappa^{\text{inc}} = 38\right] = 0.017$, $T_\kappa\left[\varphi_\kappa = 0^\circ, a_\kappa^{\text{inc}} = 38\right] = 0.866$.

The results shown in Figs. 2 - 8 are obtained by means of the iterative scheme (71). We point out some features of the numerical realisation of the algorithm (71). Figs. 9 and 10 display the number Q of iterations of the algorithm (71) that were necessary to obtain the results (analysis of scattering and generation properties of the non-linear structure) shown in Fig. 8. In Fig. 9 (left) we can see the number of iterations of the algorithm (71) for $\varphi_\kappa = 0^\circ$, the range of amplitudes $a_\kappa^{\text{inc}} \in [0, 38]$ and the range of increments $\Delta a_\kappa^{\text{inc}} = 1$. Similarly, in Fig. 9 (right), we have the following parameters: $a_\kappa^{\text{inc}} = 38$, $\varphi_\kappa \in [0^\circ, 90^\circ)$ and $\Delta\varphi_\kappa = 1^\circ$. The results

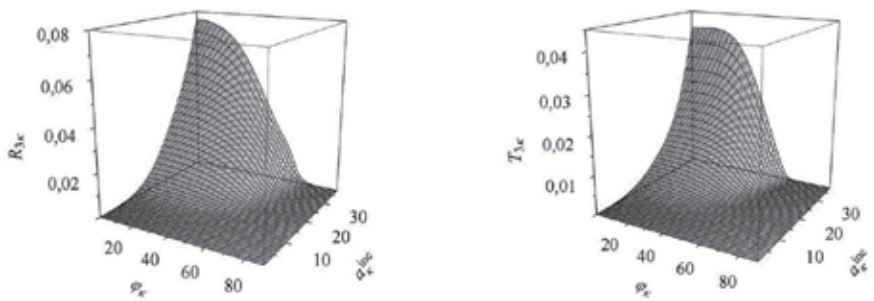


Fig. 7. Generation properties of the non-linear structure at the frequency of the third harmonic $3\kappa = 0.75$: $R_{3\kappa}\left[\varphi_\kappa, a_\kappa^{\text{inc}}\right]$ (left), $T_{3\kappa}\left[\varphi_\kappa, a_\kappa^{\text{inc}}\right]$ (right)

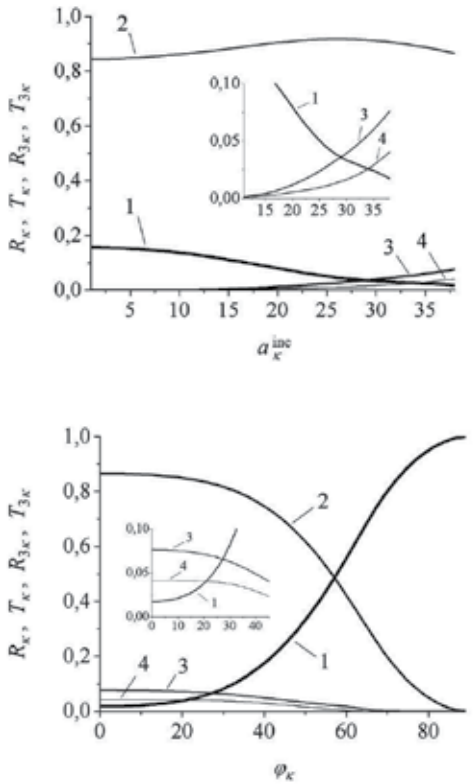


Fig. 8. Scattering and generation properties of the non-linear structure, $\kappa = 0.25$, $3\kappa = 0.75$, for $\varphi_\kappa = 0^\circ$ (left) and $a_\kappa^{\text{inc}} = 38$ (right): #1 … R_κ, #2 … T_κ, #3 … $R_{3\kappa}$, #4 … $T_{3\kappa}$

shown in Fig. 9 are also reflected in Fig. 10. Here the dependencies on the portion of the total energy generated in the third harmonic $W_{3\kappa}/W_\kappa$ are presented that characterise the iterative processes. We see that the number of iterations essentially depends on the energy generated

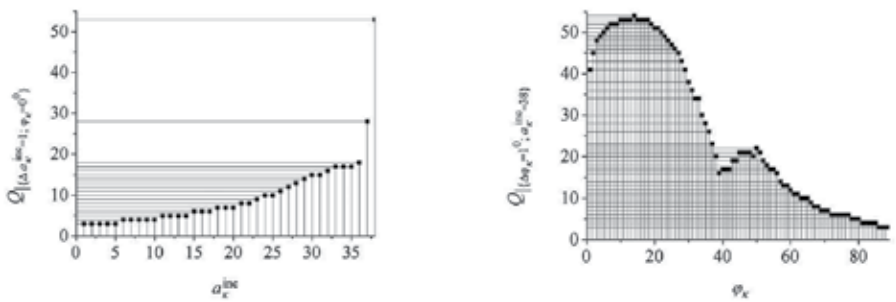


Fig. 9. The number of iterations of the algorithm in the analysis of the generating and scattering properties of the non-linear structure ($\kappa = 0.25$, $3\kappa = 0.75$): $Q|_{\{\Delta a_\kappa^{\text{inc}}=1,\ \varphi_\kappa=0^\circ\}}$ for $\Delta a_\kappa^{\text{inc}} = 1$ and $\varphi_\kappa = 0^\circ$ (left), $Q|_{\{\Delta\varphi_\kappa=1^\circ,\ a_\kappa^{\text{inc}}=38\}}$ for $\Delta\varphi_\kappa = 1^\circ$ and $a_\kappa^{\text{inc}} = 38$ (right)

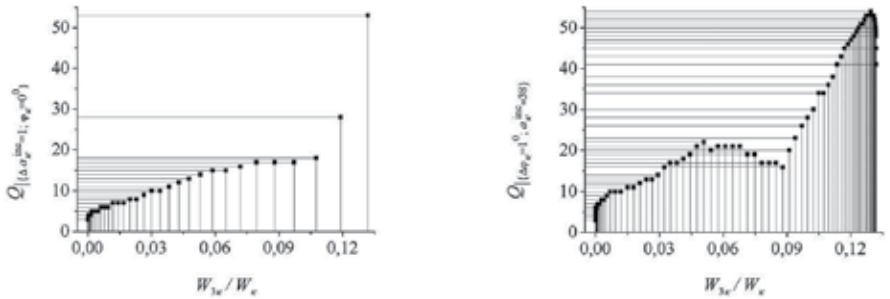


Fig. 10. The number of iterations of the algorithm in the analysis of the generating and scattering properties of the non-linear structure ($\kappa = 0.25$, $3\kappa = 0.75$) in dependence on the value $W_{3\kappa}/W_{\kappa}$: $Q|_{\{\Delta a_{\kappa}^{\mathrm{inc}}=1,\ \varphi_{\kappa}=0^{\circ}\}}$ for $\Delta a_{\kappa}^{\mathrm{inc}} = 1$ and $\varphi_{\kappa} = 0^{\circ}$ (left), $Q|_{\{\Delta\varphi_{\kappa}=1^{\circ},\ a_{\kappa}^{\mathrm{inc}}=38\}}$ for $\Delta\varphi_{\kappa} = 1^{\circ}$ and $a_{\kappa}^{\mathrm{inc}} = 38$ (right)

in the third harmonic of the field by the non-linear structure.
The numerical results presented above were obtained by the iterative scheme (71) based on Simpson's quadrature rule, see Angermann & Yatsyk (2010). In the investigated range of parameters of the non-linear problem, the dimension of the resulting system of algebraic equations was $N = 501$, the relative error of calculations did not exceed $\xi = 10^{-7}$. Finally, it should be mentioned that the analysis of the problem (31), (C1) – (C4) can be carried out by solving the system of non-linear integral equations (52) and (55) as well as by solving the non-linear boundary value problems of Sturm-Liouville type (60). The numerical investigation of the non-linear boundary value problems (60) is based on the application of the finite element method Angermann & Yatsyk (2008), Angermann & Yatsyk (2010) Samarskij & Gulin (2003).

10. Conclusion

We presented a mathematical model and numerical simulations for the problem of resonance scattering and generation of harmonics by the diffraction of an incident wave packet by a non-linear layered cubically polarised structure. This model essentially extends the model proposed earlier in Yatsyk (September 21-24, 2009), Angermann & Yatsyk (2010), where only the case of normal incidence of the wave packet has been investigated. The involvment of the condition of phase synchronism into the boundary conditions of the problem allowed us to eliminate this restriction. The incident wave packet may fall onto the non-linear layered structure under an arbitrary angle. The wave packets under consideration consist of a strong field leading to the generation of waves and of weak fields which do not lead to the generation of harmonics but have a certain influence on the process of scattering and wave generation by the non-linear structure. The research was focused on the construction of algorithms for the analysis of resonant scattering and wave generation by a cubically non-linear layered structure. Results of calculations of the scattering field of a plane wave including the effect of the third harmonic generation by the structure were given. In particular, within the framework of the closed system of boundary value problems under consideration it could be shown that the imaginary part of the dielectric permittivity, which depends on the value of the non-linear part of the polarisation at the excitation frequency, characterises the loss of energy in the non-linear medium (at the frequency of the incident field) caused by to

the generation of the electromagnetic field of the third harmonic (at the triple frequency). For a sufficiently strong excitation field, the magnitude of the total energy generated by the non-linear structure at the triple frequency reaches 13.2 % of the total energy dissipated at the frequency of excitation. In addition, the paper presented the results describing the scattering and generation properties of the non-linear layered structure.

11. References

Agranovich, V. & Ginzburg, V. (1966). *Spatial Dispersion in Crystal Optics and the Theory of Excitons*, Interscience, Innsbruck.

Akhmediev, N. & Ankevich, A. (2003). *Solitons*, Fizmatlit, Moscow.

Angermann, L. & Yatsyk, V. (2008). Numerical simulation of the diffraction of weak electromagnetic waves by a Kerr-type nonlinear dielectric layer, *Int. J. Electromagnetic Waves and Electronic Systems* 13(12): 15–30.

Angermann, L. & Yatsyk, V. (2010). Mathematical models of the analysis of processes of resonance scattering and generation of the third harmonic by the diffraction of a plane wave through a layered, cubically polarisable structure, *Int. J. Electromagnetic Waves and Electronic Systems* 15(1): 36–49. In Russian.

Butcher, P. (1965). Nonlinear optical phenomena, *Bulletin 200*, Ohio State University, Columbus.

Kleinman, D. (1962). Nonlinear dielectric polarization in optical media, *Phys. Rev.* 126(6): 1977–1979.

Kravchenko, V. & Yatsyk, V. (2007). Effects of resonant scattering of waves by layered dielectric structure with Kerr-type nonlinearity, *Int. J. Electromagnetic Waves and Electronic Systems* 12(12): 17–40.

Miloslavski, V. (2008). *Nonlinear Optics*, V.N. Karazin Kharkov National University, Kharkov.

Samarskij, A. & Gulin, A. (2003). *Chislennye metody matematicheskoi fiziki (Numerical Methods of Mathematical Physics)*, Nauchnyi Mir, Moscow. In Russian.

Schürmann, H. W., Serov, V. & Shestopalov, Y. (2001). Reflection and transmission of a TE-plane wave at a lossless nonlinear dielectric film, *Physica D* 158: 197–215.

Serov, V., Schürmann, H. & Svetogorova, E. (2004). Integral equation approach to reflection and transmission of a plane te-wave at a (linear/nonlinear) dielectric film with spatially varying permittivities, *J. Phys. A: Math. Gen.* 37: 3489–3500.

Shestopalov, V. & Sirenko, Y. (1989). *Dynamical Theory of Gratings*, Naukova, Dumka, Kiev.

Shestopalov, Y. & Yatsyk, V. (2007). Resonance scattering of electromagnetic waves by a Kerr nonlinear dielectric layer, *Radiotekhnika i Elektronika (J. of Communications Technology and Electronics)* 52(11): 1285–1300.

Shestopalov, Y. & Yatsyk, V. (2010). Diffraction of electromagnetic waves by a layer filled with a Kerr-type nonlinear medium, *J. of Nonlinear Math. Physics* . 17(3): 311-335.

Sirenko, Y., Shestopalov, V. & Yatsyk, V. (1985). Elements of the spectral theory of gratings, *Preprint 266*, IRE NAS Ukraine, Kharkov.

Smirnov, V. (1981). *Course of Higher Mathematics, Vol. 4, Ch. 2*, Nauka, Moscow.

Smirnov, Y., Schürmann, H. & Shestopalov, Y. (2005). Propagation of TE-waves in cylindrical nonlinear dielectric waveguides, *Physical Review E* 71: 0166141–10.

Vainstein, L. (1988). *Electromagnetic Waves*, Radio i Svyas, Moscow. In Russian.

Vinogradova, M., Rudenko, O. & Sukhorukov, A. (1990). *Wave Theory*, Nauka, Moscow.

Yatsyk, V. (2006). Diffraction by a layer and layered structure with positive and negative susceptibilities of Kerr-nonlinear media, *Usp. Sovr. Radioelektroniki* 8: 68–80.

Yatsyk, V. (2007). About a problem of diffraction on transverse non-homogeneous dielectric layer of Kerr-like nonlinearity, *Int. J. Electromagnetic Waves and Electronic Systems* 12(1): 59–69.

Yatsyk, V. (June 21-26, 2010). Generation of the third harmonic at the diffraction on nonlinear layered structure, *International Kharkov Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW-2010)*, Kharkov, Ukraine, pp. A–20:1–3.

Yatsyk, V. (September 21-24, 2009). Problem of diffraction on nonlinear dielectric layered structure. Generation of the third harmonic, *International Seminar/Workshop on Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory (DIPED-2009)*, IAPMM, NASU, Lviv, Ukraine, pp. 92–98.

9

Numerical Modeling of Reflector Antennas

Oleg A. Yurtcev and Yuri Y. Bobkov
Belarusian state university of informatics and radioelectronics
Belarus

1. Introduction

The numerical modeling of reflector antennas is a necessary stage of their design. Due to numerical modeling dimensions of all antenna elements are defined. The more factors are accounted during antenna numerical modeling the more accurately the antenna elements dimensions are defined. There are many methods used in the programs of antenna numerical modeling: geometric optics method; aperture method; geometric theory method of diffraction; physical optics method, integral equations method; finite elements method. By now there are many papers in which the different aspects of reflector antenna numerical modeling are discussed. For determination of the field antenna reflector in regions of main lobe and first side lobes in front semi-space the aperture method is used; for determination of the field in full semi-space the physical optics method is used (Chen & Xu, 1990; Charles, 1975; Rusch, 1974). The geometric theory of diffraction (Narasimhan & Govind, 1991; Rahmat-Samii, 1986; Narasimhan et al, 1981) and moment method (Khayatian & Rahmat-Samii, 1999) are used for determination of the field in back semi-space, for determination of field features in front semi-space related with diffraction of the field on the edge of paraboloid and hyperboloid surfaces and for modeling the feed-horn. In a number of papers different approaches are used for simplification of analytical expressions for calculation of antenna fields to reduce a mathematical model of antenna and to simplify modeling program (Rahmat-Samii, 1987). A number of works deal with research into the field in near-field zone (Narasimhan & Christopher, 1984; Fitzgerald, 1972; Houshmand et al., 1988; Watson, 1964). But the results are not reduced to numerical data in that volume which is necessary for antenna design. The field distribution in near-field zone is described in detail for plane aperture at uniform its excitation (Laybros et al., 2005), but for reflector antennas such research was not provided. The reflector antenna in receiving mode is not discussed in literature, however at designing antenna for radioimaging systems it is necessary to know of field distribution in the focal region at receiving of the wave from near-field zone points. The issue of isolation of channels in multi-beam reflector antenna at receiving of the wave from near-field zone is not analyzed too. Without analysis of the isolation between channels it is impossible to analyze the quality of imaging in radioimaging systems.
In literature a number of works deal with describing the feed-horns in monopulse reflector antennas (Hannan, 1961; Scolnic, 1970). There is a little information on numerical characteristics description the regularity in monopulse reflector antenna.
In the present chapter the mathematical model of the single-reflector paraboloid antenna and double-reflector paraboloid Cassegrain antenna is based on physical optics method

with the same features in comparison with frequently used models. These features are the following:
a) the feed-horn in the form of the pyramidal horn is not accurate; a limited feed-horn aperture dimensions, its depth and influence of these dimensions on distribution of the amplitude and phase of the field on the horn aperture are assumed; the feed-horn field is determined based on amplitude and phase of the field on the aperture by Kirchgoff integral;
b) it is supposed that paraboloid in single-reflector and hyperboloid in double-reflector antenna are located on the prefixed distance from feed-horn aperture plane (the approximation of the far-field zone is not used);
c) paraboloid in the double-reflector antenna is located on the unknown distance from hyperboloid (the approximation of the far-field zone is used);
f) in radiation mode, the point, in which the field is defined, is located in any zone of space (far-field, intermediate, near-field); in receiving mode the point of spherical wave source is located also in any space zone.
The geometrical theory of diffraction is used only for analysis of the field in back semi space, but in the chapter the analysis results are not present. Using of Kirchgoff integral for calculation of the field of feed-horn (in radiation mode), waveguide excitation theory (in receiving mode) and physical optics method at determination of the field of hyperboloid and paraboloid allow to avoid limitations on wave dimensions of the paraboloid. Simulation time of problem and needed memory value of computer is less than for universal electromagnetic simulation programs such as CST MICROWAVE STUDIO, HFSS, FEKO. The modeling accuracy is about the same.

2. Mathematical model of reflector antenna in radiation and receiving modes

2.1 Antenna geometry

The single-reflector and double-reflector Casserrain antenna (reflector is parabolic, sub-reflector is hyperbolic) are analyzed in this work. The double-reflector antenna within coordinate system X,Y,Z and its geometric dimensions are shown in figure 1. The antenna elements involved are: 1 – paraboloid; 2 – hyperboloid; 3 – feed horn; 4 – rectangular waveguide. Antenna element dimensions and markings are the following: F_h – phased center of feed horn; F –parabolic focus; D_p – parabolic diameter; D_g – hyperboloid diameter; F_P – parabolic focus distance; F_{g1} – far focus hyperboloid distance; F_{g2} – near focus hyperboloid distance; $2\theta_{M\,max}$ $2\Theta_0$ – parabolic aperture angle; M – point on parabolic surface; R_M, θ_M, φ_M R_M, θ_M, φ_M – spherical coordinates of M point with respect to parabolic focus F.
The space point P is shown in fig2. It is in this point that the field is determined in this point in radiation mode. In receiving mode the EM-field source is located in point P. The position of point P is set by spherical coordinates R, θ, φ. The projection of P point on XY-plane is shown in figure 2 as Pxy point with coordinates R, θ, φ.
The combination of physical optics method (PO) and geometric theory of diffraction (GTD) are used in mathematical model of reflector antenna in the radiation mode. The physical optics method is used for calculation of field in front semi-space (θ<90°). The GTD method is used in mathematical model of reflector antenna for calculation field in back semi-space (θ>90°). The point P is located only in front semi-space for receiving mode and antenna is analyzed by physical optics method. The theory of waveguide excitation is used for calculation of power level on waveguide input.

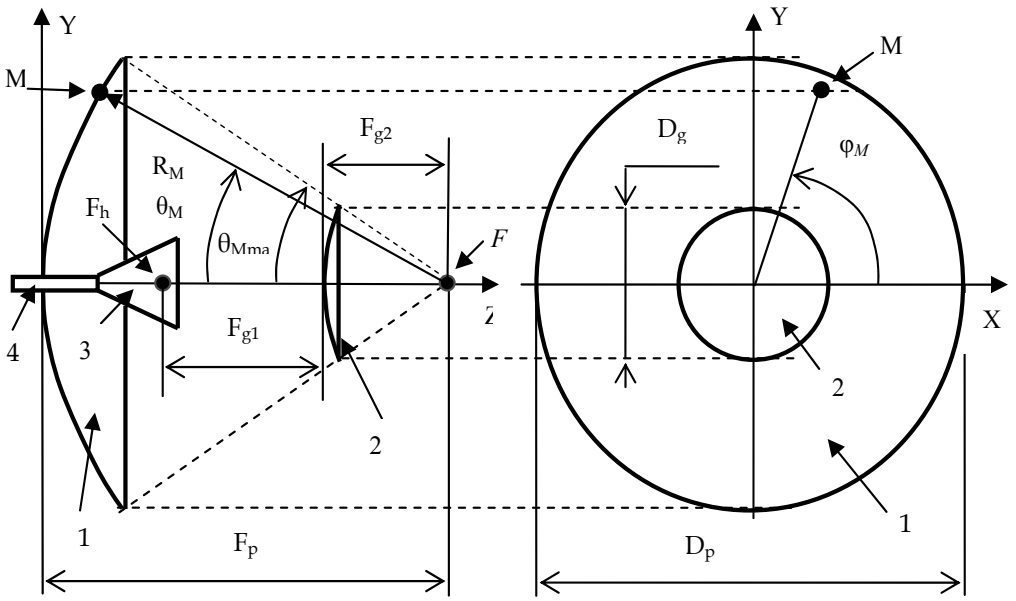


Fig. 1. Double-reflector antenna

2.2 Single-reflector antenna in radiation mode

Pyramidal horn is used as feed-horn. The feed horn is executed by rectangular waveguide – figure 3. The horn dimensions A_h, B_h are aperture dimensions, R_h is horn depth. The case when polarization plane of feed horn coincides with YZ plane is considered. The dimension of waveguide cross section satisfy a uniqueness condition of wave TE_{10} : $a<\lambda<2a$ $a < \lambda < 2a$, where λ – wavelength.

The mathematical model includes the following known equations.

The complex amplitude of field in a rectangular waveguide at horn input:

$$\dot{E}_y = E_m \cos(\pi x / a) \tag{1}$$

where a- is the dimension of the wide wall of rectangular waveguide;

$E_m = \sqrt{Ps \cdot Z/(a \cdot b)}$ - is electric field amplitude in the center of side «a» of the waveguide.

$Z = 120\pi / \sqrt{\varepsilon[1-(\lambda/2a)^2]}$ - is characteristic impedance of the waveguide;

ε - is related permittivity of waveguide internal environment.

Further an approximation is used: the wave in feed horn has spherical wave front. The wave source is located in horn vertex – in point O in figure 3. In this wave the field phase Ψ along the direction R_{oq} from horn vortex to Q point on aperture Sh is changed according to the law: $\Psi = 2\pi R_{oq} / \lambda$. The field amplitude is changed proportionally $1/ R_{oq}$. As a result of it the distributions of field phases $\Psi_s(x,y)$ and field amplitudes $E_s(x,y)$ on the horn aperture are calculated by expressions:

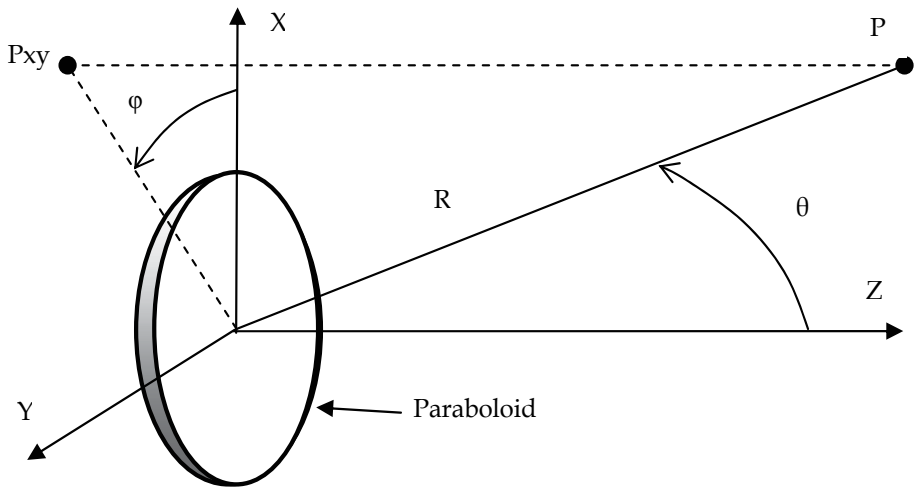


Fig. 2. Antenna and point P in space.

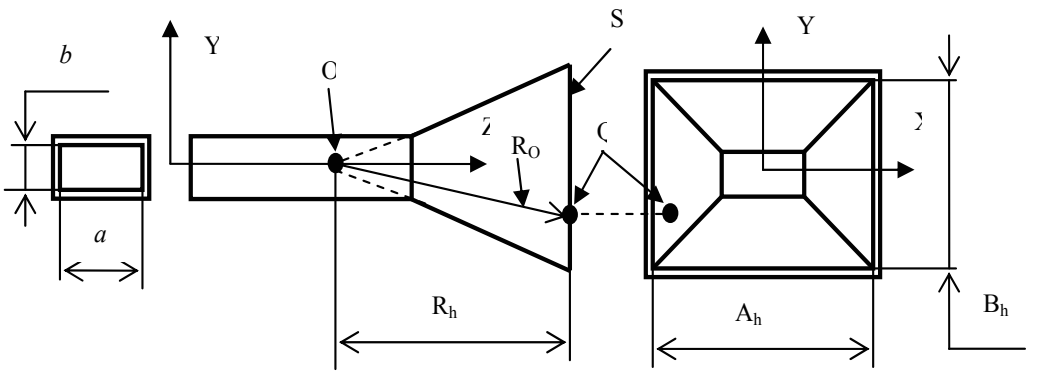


Fig. 3. Rectangular waveguide and feed horn

$$\Psi_s(x,y) = -2\pi \cdot [R_{oq}(x,y) - R_{oq}(0,0)]/\lambda ;$$
$$E_s(x,y) = E_m R_{oq}(0,0)\cos(\pi x/A_h)/R_{oq}(x,y) \quad (2)$$

where E_m is amplitude of field in the center of wide waveguide wall;

$$R_{oq}(x.y) = \sqrt{R_h^2 + x^2 + y^2} \; ; \; -0{,}5A_h \le x \le 0{,}5A_h ; \quad -0{,}5B_h \le y \le 0{,}5B_h \quad (3)$$

The field on paraboloid surface is calculated by Kirchhoff integral according to the field on aperture. The field is calculated in arbitrary point M having rectangular coordinates X_M, Y_M, Z_M and spherical coordinates R_M, θ_M, φ_M (see fig.1).

$$\vec{E}_M \approx i\frac{1}{2\lambda}\int_{S_h} E_s\left[\vec{\theta}_o(\eta\cos\theta_M+1)\cos\ _M-\vec{\varphi}\ _M(\cos\theta_M+\eta)\sin\ _M\right]\frac{\exp(i\Psi_s-ikR_{QM})}{R_{QM}}dS \quad (4)$$

where E_S, Ψ_S are determined by equations (2); $\eta=\sqrt{1-(\lambda/2Ah)^2}$; $i=\sqrt{-1}$; k=2π/λ; $\vec{\theta}_o, \vec{\varphi}_o$ are unit vectors of spherical coordinate system R_M, θ_M, φ_M; R_{QM} is the distance from point Q on horn aperture (see fig.3) to point M on the paraboloid surface. This distance is expressed in terms of rectangular coordinates of Q, M points in the coordinate system, with the geginning being in the point of paraboloid vortex O_p (see fig.4). The center of feed-horn aperture feed-horn (point Q_S) is shifted from paraboloid focus point (point F) at coordinates X,Y,Z on values D_{hx}, D_{hy}, D_{hz}. This would provides an antenna focusing in well known point P of any space zone.

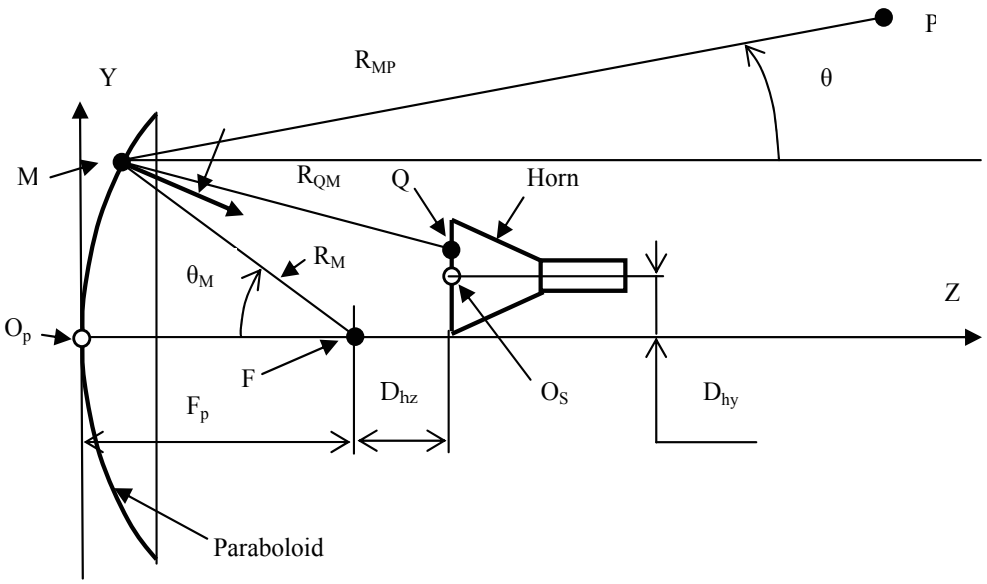


Fig. 4. Single-reflector antenna with feed horn

The expression for R_{QM} results from figure 4 by means of rectangular coordinates of the point Q and M in X,Y,Z coordinate system.

$$R_{QM}=\sqrt{(x_Q-x_M)^2+(y_Q-y_M)^2+(z_Q-z_M)^2}\ ,$$
$$x_Q=D_{hx}+x\ ;\ \ y_Q=D_{hy}+y\ ;\ \ z_Q=D_{hz}+z \quad (5)$$

$$x_M=2F_p\sin\theta_M\cos\varphi_M/(1+\cos\theta_M)\ ;\ y_M=2F_p\sin\theta_M\sin\varphi_M/(1+\cos\theta_M)\ ;$$
$$z_M=F_p\left[1-2\cos\theta_M/(1+\cos\theta_M)\right] \quad (6)$$

The angles θ_M, φ_M are calculated by coordinates x_M, y_M, x_Q, y_Q using relations:

$$\sin\theta_M = \sqrt{(x_M - x_Q)^2 + (y_M - y_Q)^2}\,/R_{QM};$$
$$tg\varphi_M = (x_M - x_Q)/(y_M - y_Q) \tag{7}$$

The point M on the paraboloid surface is the point of crossing of two line systems, which are the paraboloid surface lying in XZ and YZ planes. The distance between lines on coordinates X and Y are marked as ΔX and ΔY. The values of ΔX and ΔY are selected according to the criterion of convergence of the calculations of side lobe levels and antenna gain, i.e. by parameters which are most critical to ΔX/λ and ΔY/λ values. The results of numerical modeling show that the ΔX/λ>3 and ΔY/λ>3 are sufficient.

The vector of the magnetic field $\vec{H}_M$ in point M and than the vector of surface current density are determined in terms of the field $\vec{E}_M$. In conformity with PO method:

$$\vec{J}_s = 2[\vec{n}_o, \vec{H}_M] = \vec{J}_x + \vec{J}_y + \vec{J}_z \tag{9}$$

where $\vec{J}_x, \vec{J}_y, \vec{J}_z$ are components of $\vec{J}_s$ vector in rectangular coordinate system depending on M point coordinates x_M, y_M, z_M; the $\vec{n}_o$ is a unit vector perpendicular to paraboloid surface in point M.

The vector potential $\vec{A}$ method is used for determination of $\vec{E}$ field of the currents $\vec{J}_x, \vec{J}_y, \vec{J}_z$ in the point P (see figures 2, 4):

$$\vec{E} \approx -i\frac{60\pi}{\lambda}\vec{A} = \vec{E}_x + \vec{E}y + \vec{E}_z; \quad \vec{A} = \int_{S_p} \vec{J}_s \frac{\exp(-ikR_{MP})}{R_{MP}} dS;$$
$$R_{MP} = \sqrt{(x_M - x_P)^2 + (y_M - y_P)^2 + (z_M - z_P)^2} \tag{10}$$

where S_p are paraboloid surfaces.

The rectangular coordinates of P point x_p, y_p, z_p are associated with spherical coordinates R, θ, φ as:

$$x_P = R\sin\theta\cos\varphi;\ y_P = R\sin\theta\sin\varphi;\ z_P = R\cos\theta \tag{11}$$

The vector $\vec{E}$ projections on unit vectors $\vec{\theta}_o$ и $\vec{\varphi}_o$ are determined by expressions:

$$E_\theta = (E_x\cos\varphi + E_y\sin\varphi)\cos\theta - E_z\sin\theta;$$
$$E_\varphi = (-E_x\sin\varphi + E_y\cos\varphi)\cos\theta \tag{12}$$

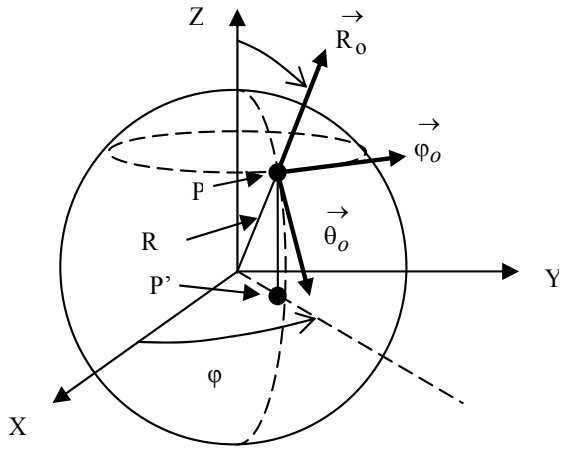


Fig. 5. Spherical coordinates system

The antenna directivity diagrams $F_\theta(\theta,\varphi)$, $F_\varphi(\theta,\varphi)$ and antenna gain G are calculated based on E_θ и E_φ components.

$$G = E_{max}^2 R^2 / 60 P_s \tag{13}$$

Where E_{max} is the maximum value of the electric field amplitude on sphere R=const; P is radiation power.

2.3 A single-reflector antenna in receiving mode.

A reflector antenna can receive a spherical wave from any points of far, intermediate and near field zones space. Let a spherical wave source be located in the point P shown in fig. 4. The amplitude of the wave electric field of is equal E_i. It is necessary to calculate power P_i entering the waveguide. The algorithm of power P_i calculation includes the following steps:

Step 1. The vector of surface current on paraboloid surface is calculated based on the spherical wave magnetic field $\vec{H}_i$ and the wave propagation direction using the formula similar to (9).

Step 2. The field $\vec{E}_S$ in the point Q(x_Q, y_Q, z_Q) on feed horn aperture is calculated by components $\vec{J}_x, \vec{J}_y, \vec{J}_z$ of surface current $\vec{j}_s$ using the formula similar to (10)

$$\vec{E}_S \approx -i\frac{60\pi}{\lambda}\vec{A} = \vec{E}_x + \vec{E}y + \vec{E}_z;$$
$$\vec{A} = \int_{S_p} \vec{J}_s \frac{\exp(-ikR_{MQ})}{R_{MQ}} dS; \tag{14}$$
$$R_{MQ} = \sqrt{(x_M - x_Q)^2 + (y_M - y_Q)^2 + (z_M - z_Q)^2}$$

Step 3. The amplitude of TE_{10} wave in the rectangular waveguide is defined by $\vec{E}_S$ field. This problem is solved by the own wave method using the waveguide excitation theory.

In conformity with this theory the $\vec{E}$ and $\vec{H}$ field in waveguide is presented as the sum of the own waveguide waves $\vec{E}_\nu$ and $\vec{H}_\nu$, where ν is generalized index describing a wave type and its propagation direction:

$$\vec{E} = \sum_{(\nu)} C_\nu \vec{E}_\nu , \quad \vec{H} = \sum_{(\nu)} C_\nu \vec{H}_\nu \tag{15}$$

where C_ν is an excitation coefficient related to off-site sources – the density of off-site electric current $\vec{J}_e$ and magnetic $\vec{J}_h$ currents:

$$C_\nu = \frac{1}{N_\nu} \int_V \left[\vec{J}_e \vec{E}_{-\nu} - \vec{J}_h \vec{H}_{-\nu} \right] dV \tag{16}$$

where V is the volume in which the off-site sources of the field are located; N_ν is the norm, given by

$$N_\nu = \int_S \{ [\vec{E}_\nu, \vec{H}_{-\nu}] - [\vec{E}_{-\nu}, \vec{H}_\nu] \} \vec{m}_o dS \tag{17}$$

In equations (14)-(16) the $\vec{E}_\nu, \vec{H}_\nu$ are advanced own waves; $\vec{E}_{-\nu}, \vec{H}_{-\nu}$ are reversed own waves; S is a waveguide cross-section area; $\vec{m}_o$ is a unit vector perpendicular to the waveguide cross-section.

The equations concerned are used for solution of problems of TE_{10} wave excitation in the rectangular waveguide with cross-section dimensions A_h and B_h without accounting transformation of the waveguide to aperture horn. It is assumed that other waves except TE_{10} cannot propagate. The integration in (15) is carried out on the horn aperture. On the horn aperture the vector $\vec{J}_e$=0, and vector $\vec{J}_h$ is expressed by the field $\vec{E}_S$ component tangent to horn aperture. The axis of excited waveguide is oriented along the Z-axis. In this case the $\vec{m}_o = \vec{z}_o$, where $\vec{z}_o$ is a unit vector parallel to the Z-axis. The vector $\vec{J}_h$ is expressed by the vector $\vec{E}_S$

$$\vec{J}_h = -[\vec{z}_o, \vec{E}_S] \tag{17}$$

Using current formulas for vector components of electrical and magnetic fields of X TE_{10} wave, it's not difficult to deduce a formula for the norm of this wave:

$$N\nu = N_{h10} = \frac{E_m^2}{Z_v} A_h \cdot B_h \tag{18}$$

Where $Z_v = \sqrt{\frac{\mu_a}{\varepsilon_a}} / \sqrt{1 - \left(\frac{\lambda}{2A_h}\right)^2}$ is characteristic cross-section impedance of the waveguide with Ar and Br dimensions for TE_{10} wave; ε_a, μ_a are absolute permittivity and absolute permeability of the medium filling the cave of the waveguide; E_m is the amplitude of the TE_{10} wave electrical field in the center of the waveguide cross-section.
After simple manipulations the formula for the TE_{10} wave electrical field is as follows, (there is only one E_y component for the vector $\vec{E}$)

$$E = E_y = \frac{\cos(\pi x / A_h)}{A_h \cdot B_h} \int_{x=-0,5A_h}^{0,5A_h} \int_{y=-0,5B_h}^{0,5B_h} E_{sy}(x,y) \cos(\pi x / A_h) dx \cdot dy \tag{19}$$

where $E_{sy}(x,y)$ is a component of an off-set field (the field of paraboloid) at the horn aperture which is tangent to the horn aperture and parallel to the Y-axis.
The expression for field amplitude E_{max} in the center of the wide side of the horn for power Pr received by horn results from (19):

$$E_{max} = \frac{1}{A_h \cdot B_h} \left| \int_{x=-0,5A_h}^{0,5A_h} \int_{y=-0,5B_h}^{0,5B_h} E_{py}(x,y) \cos(\pi x / A_h) dx \cdot dy \right| \tag{20}$$

$$Pr = E_{max}^2 A_h \cdot B_h / 4Z_v \tag{21}$$

2.4 Double-reflector Cassegrain antenna in radiation mode

In fig.6 a paraboloid (1) and a hyperboloid (2) with additional designations are shown, Oq – the apex of the paraboloid (1); Oq is the apex of the hyperboloid; F1 and F2 are the near and the far foci, N is a point on the hyperboloid surface; R_1 is a distance between F_1 and N points, R_2 is a distance between F_2 and N points. Focus F of the paraboloid and the nearest focus of the hyperboloid coincide. The distant focus of the hyperboloid is agreed with the phase center of the horn. To focus the antenna on the given distance we move the feed-horn – hyperboloid system along the z-axis by Dz distance, to scan – we rotate the hyperboloid around Og point. In figure 6 one of F1-N-M rays is shown as a chain-line.

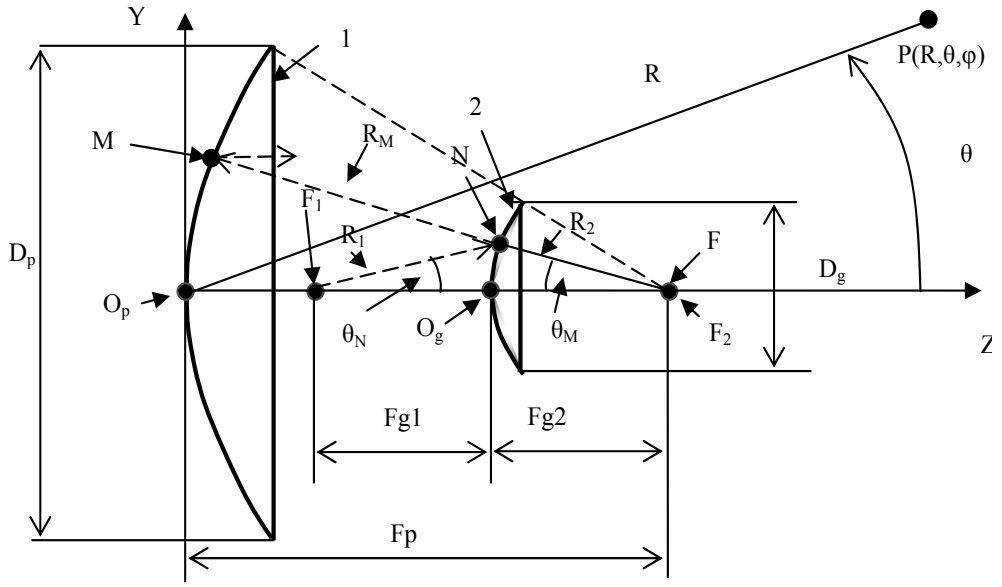


Fig. 6. Paraboloid-hyperboloid system

The configuration of the hyperboloid may be described with the following formulas:

$$R_1 = F_{g2}(1+e)/|1-e\cdot\cos\theta_N|; \tag{22}$$

$$tg(0{,}5\theta_M) = (e+1)/(e-1)\cdot tg(0{,}5\cdot\theta_N) \tag{23}$$

where e- is the eccentricity of the hyperboloid ($e \approx 1{,}2\ldots2$)

$$e = (F_{g1}/F_{g2}+1)/(F_{g1}/F_{g2}-1) \tag{24}$$

The sequence of antenna field calculation in the point P(R,θ,φ) is the following:

1. Using formula (14) makes it possible to determine the electrical field vector $\vec{E}_N(X_N, Y_N, Z_N)$ on the surface of the hyperboloid in the point N and then to calculate the vector of the magnetic field strength $\vec{H}_N(X_N, Y_N, Z_N)$. In formula (4) we substitute θ_M angle for θ_N angle, R_{QM} for R_{QN}. The angle θ_N is changed in limits

$$0 \le \theta_N \le \theta_{N\max}. \tag{23}$$

The $\theta_{N\max}$ and $\theta_{M\max}$ angles are both performed in formula (23), R_{QN}, is defined by formula 5 with substituting R_M, θ_M, φ_M coordinates of point M for coordinates of point N in formula (5):

$$X_N = R_1\cos\theta_N\cos\varphi_N;\ \ Y_N = R_1\cos\theta_N\sin\varphi_N;$$
$$Z_N = F_p - (F_{g1}+F_{g2}) + R_1\cos\theta_N, \tag{24}$$

where φ_N – an angular coordinate of point N in the XY plane

2. Using the formula which is analogous to formula (9) gives a vector of current density on the hyperboloid surface

$$\vec{J}_{sN} = 2[\vec{n}_o, \vec{H}_N] = \vec{J}_{Nx} + \vec{J}_{Ny} + \vec{J}_{Nz} \tag{25}$$

3. Knowing $\vec{J}_{sN}$ current we determine the field on the surface of the paraboloid in point M. We use the formula given by (10)

$$\vec{E}_M \approx -i\frac{60\pi}{\lambda}\vec{A}; \quad \vec{A} = \int_{S_g} \vec{J}_{sN} \frac{\exp(-ikR_{NM})}{R_{NM}} dS;$$
$$R_{NM} = \sqrt{(x_N - x_M)^2 + (y_N - y_M)^2 + (z_N - z_M)^2}, \tag{26}$$

Where S_g is a paraboloid surface.

4. Using the electrical field vector $\vec{E}_M$ we determine the magnetic field vector $\vec{H}_M$ in the point M and then we calculate the current on the surface of the paraboloid and the field in the point P according to formulas (9)-(12). The formulas involved are used in the program for the numerical simulation of different types of reflector antennas as well as for researching field characteristics in the near-field region

3. Results of numerical simulation and its discussion

3.1 The field distribution in the near-field zone in radiation mode

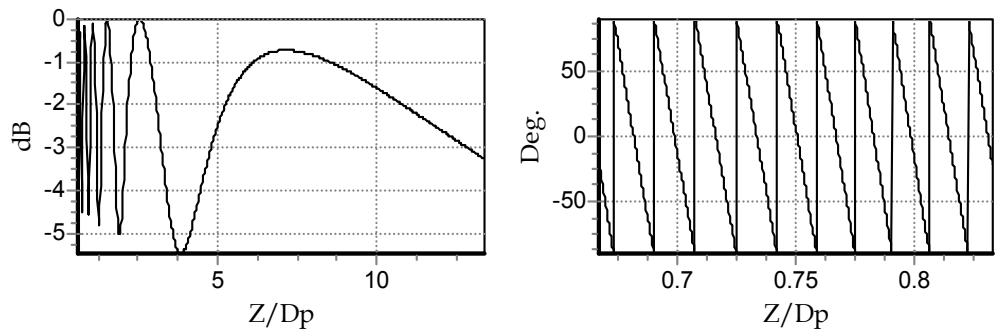
The simulation was made for antennas with the paraboloid diameter D_p =(10…100)λ and different ratio F_p / D_p. To demonstrate the main principles we took a single reflector antenna with $Dp = 30\lambda$ and F_p / D_p =0.5 as an example and studied field distribution in the tangent plane (z=const), along the focal axis (z – direction) and depending on the angle θ. The calculations were made for the near-field zone, the intermediate zone and the far-field zone focusing the antenna into the far-field zone and into the given point of the near-field zone. We considered the technology of scanning during focusing the antenna. The sizes of the feed-horn Ah, Bh have been chosen to bring the illumination level of the edge of the reflector with respect to its center in the E and H planes to about 0.3. It coresponds to the maximum antenna gain. All numerical results are given for the plane E.

The distribution of amplitudes and field-phases along the focal axis (Z-distribution) in the near zone is shown in fig.7. The coordinate Z is dependent on the focus point. The antenna is focused on the far-field zone.

It is shown in fig.7 that moving the observation point away from reflector the field amplitude oscillates. Monotonous decrease of the field amplitude begins in the point Z_o. The value of Z_o and the depth of oscillations increase with the rise of D_p / λ. The reason for the oscillations is the interference of different Fresnel zones at the reflector aperture.

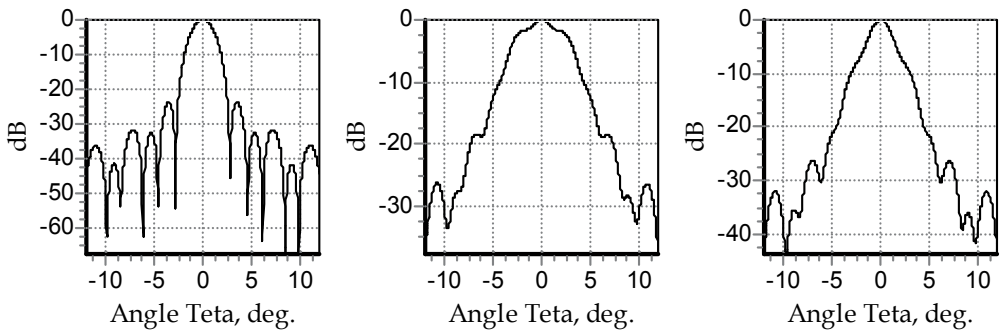
The distribution of the field phase along the focal axis is linear (fig. 7b). It indicates that the traveling field wave propagates along focal axis.

The same situation is observed in the back semi-space, but the oscillations have a less depth and as the distance from the apex of the paraboloid grows the amplitude decreases considerably faster than it happens in the front semi-space.



a) Distribution $|E(Z)|/E\max$ b) Distribution $\arg E(Z)$

Fig. 7. The field distribution in focal axis direction.



a) R=200 D_P b) R=2,5 D_P c) R=7,5 D_P

Fig. 8. The antennas field distribution on the sphere R=const.

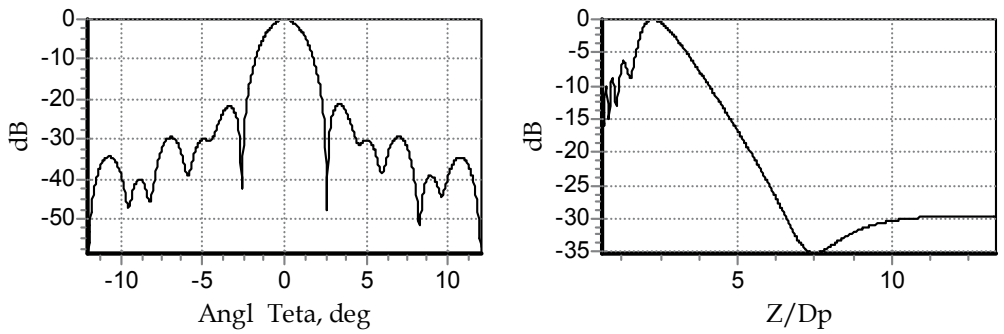
Neither of amplitude maximum points in figure 7a are an antenna's focusing points. Under a focusing point we mean the point on the sphere, on which the characteristics of distribution of the field amplitude in relation to the angle θ are close to the antenna diagram in the far zone. To illustrate this in fig.8 field distribution on the surface of the sphere R=const for the antenna focused into the far-field zone for distance a) R=200 Dp (antenna diagram); b) R=2.5 Dp; c) R=7.5 Dp (the last two amplitude maximum points are in fig.7)

For comparison in fig.9 it is shown:

the field distribution amplitude on the sphere R=2.5Dp in depending on the angle θ focused on the distance equal to the radius by shifting the feed-horn along the focal axis on D_{hz} =1,5λ (see figure 9 a);

the field amplitude distribution along the focal axis during such shifting of the feed horn (see figure 9 b)

As can be seen from fig.9 b, the field amplitude considerably increases in the focusing point that field distribution depending on the angle θ on the sphere of the antenna, R=2.5Dp focused on this distance is close to the antenna diagram in the far–field zone.



a) Distribution of field on sphere b) Distribution of field along focal axis

Fig. 9. Field distribution on sphere R=2.5 Dp and along focal axis in antenna focusing case.

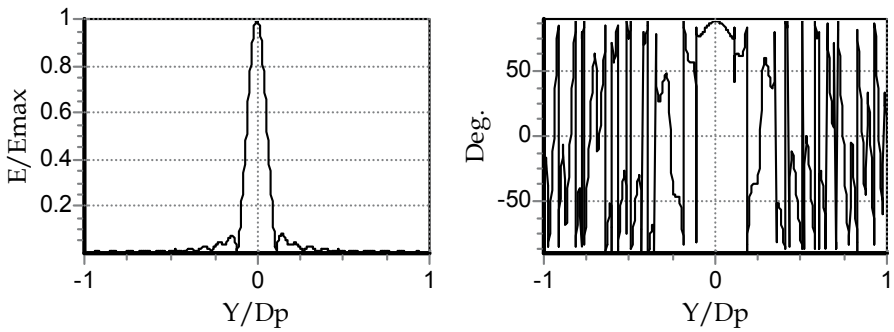


Fig. 10. The amplitude and the field phase in the focusing plane Z=const.

The dependence of the amplitude normalized to the amplitude maximum (E/E_{max}) and the field phase from the coordinate Y in the focusing plane Z=2,5 D_P is shown in fig.10. The value of Y varied within two paraboloid diameter limits ($-D_P \le Y \le D_p$).

The region in the y-direction where $E/E_{max} > 1/\sqrt{2}$ will be called the focusing zone and it will be indicated as $\Delta Y_{0,5}$. The value of $\Delta Y_{0,5}$ increases with growing distance to the focusing point and it's linearly related to the diameter of paraboloid. Analogous patterns take place in a double-reflector antenna.

The field distribution in the near-field zone in a spherical wave receiving mode.

The field distribution in the focal region of an antenna during receiving a spherical wave coming from the near-zone is of interest under optimization of the feed-horn position (or several feed-horns in a multi-beam antenna). Further the patterns are demonstrated by the example of an antenna with parameters $D_p = 30\lambda$, F_p / D_p =0,5.

The fig.11 show the field distribution along the focal axis during receiving a spherical wave coming from the point situated a) in the far zone ($R = 200D_p$) b) in the near field zone ($R = 2D_p$). The coordinate Z is counted out from paraboloid apex (point O_P in fig. 6.).

At decreasing distance to the focusing point the width of the region on the focal plane occupied by the main lobe is increasing (fig.11).
It's seen that with fig. 11a the field maximum is located in the paraboloid focus, but with figure 11 b the maximum is moved away paraboloid apex from the reflector on 1.41Fp. If the phase center of the feed-horn is placed in that point, the antenna will be focused on R=2Dp.
Fig. 12a depicts the diagram of an antenna focused on the far-field zone at the distance of R=2Dp. It's obvious that differences are discovered only in the side lobe region.
The diagrams of the antenna focused at distance R=200Dp into radiation mode (solid) and receiving mode (Dot) is depicted in fig. 12b. The differences in side lobe region result from different calculation methods, described in the mathematical model.

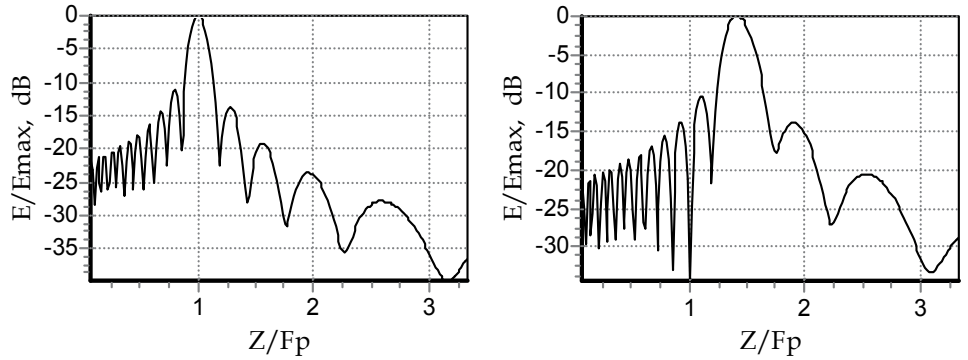
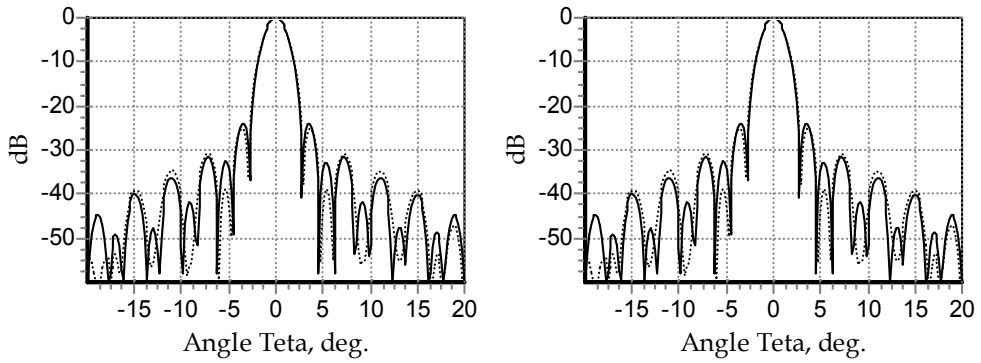


Fig. 11. The field amplitude distribution along the focal axis.



a) receiving mode:
solid - $R = 200D_p$;
dot - $R = 2D_p$

b) Far-field zone:
dot - receiving mode;
solid – radiation mode

Fig. 12. Antenna diagrams

Scanning in a single-reflector antenna. Multi-beam reflector antenna.

Scanning is produced by moving a feed-horn in a plane Z=const and it is used at the antenna focusing in far-field or near-field zones. Further peculiarities of scanning process at antenna focusing in far-field and near-field zones and features of isolation between channels

in a multi-beam antenna in receiving mode of spherical wave from near-field zone point are considered.

Regularities of scanning are demonstrated by the example of the antenna with the following parameters: D_p =300 mm; F_p =150 mm; f=37 GHz (D_p /λ=37).

The feed horn size Ar, Br are made to be less optimal according to the criteria of antenna gain maximum. This conforms to paraboloid edge illumination level on 10 dB less than in the paraboloid center. The optimal horn sizes for ratio F_p / D_p =0,5 at frequency 37 GHz are the following: A_h =9 mm, B_h =6 mm, R_h =30 mm. The diagrams of the antenna that is focused in the far-field zone at the distance R=200 Dp with three values D_{hy} =0; 20 mm; 40 mm of horn shifting of in the focal plane along Y-axis (see fig.4) are shown in fig. 13. With increasing D_{hy} the main lobe is shifted from the focal axis by the angle Q_m, the beam width $2\theta_{0,5}$ and side lobes level F_{bm} are increased too. These regularities are well-known for far field zones. These regularities remain when the antenna is focused into the near-field zone, but they are quantitatively less expressed. The diagrams with the same parameters D_{hy} for the antenna focused into neat-field zone at the distance R=4 D_P are depicted in fig. 14. The antenna focusing into this distance is produced by shifting the feed-horn along the focal axis at the distance D_{hz} =22 mm (approximately 3λ).

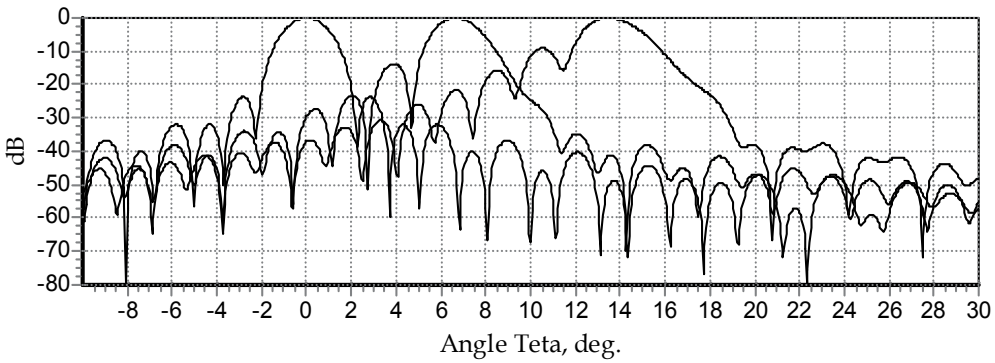


Fig. 13. The antenna diagrams during scanning. The antenna is focused into the far-field zone.

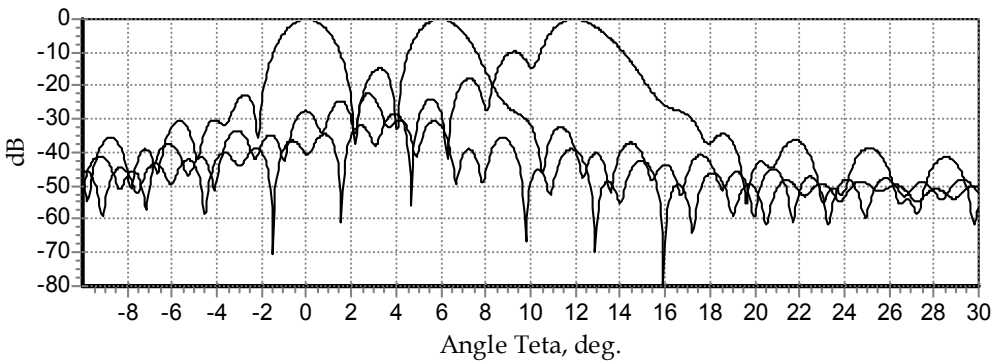


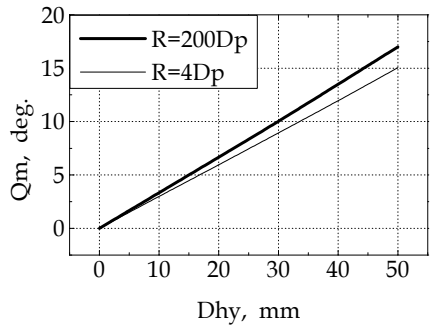
Fig. 14. The antenna diagrams during scanning. The antenna is focused into the near-field zone at the distance R=4 D_P .

The differences of antenna diagrams during scanning and focusing in far-field and near-field zones are illustrated by the following figures.

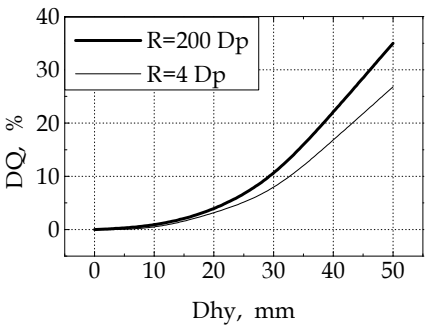
In fig. 15 the dependence of main lobe deviation angle from the focal axis (Q_m) (15-a) and the dependence of widening the main lobe (DQ%) from shifting the feed horn along Y-axis at focusing into the far-field and near-field zones (15b) shown in fig. 15b. The coefficient of widening the main lobe (DQ %) is determined from equation $DQ\%=[100\lfloor 2\theta_{0,5}(0)-2\theta_{0,5}\rfloor/2\theta_{0,5}(0)$, where $2\theta_{0,5}(0)$ is the width of the not shifted main lobe.

From figure 15 it follows, that the angle deviation of the main lobe from the focal axis and the coefficient of widening are reduced with reducing distance to the antenna focusing point.

The dependence of the coefficient of increasing of side lobe levels (DF) and reducing antenna gain (DG) versus shifting the horn along Y-axis at focusing of antenna in far-field and near-field zones are shown in figure 16. The values DF and DG are given from DF=Fbm-Fbm(0), where Fbm(0) is a side lobe level with the not deviated main lobe; DG=G(0)-G, where G(0) is antenna gain with not deviated main lobe.

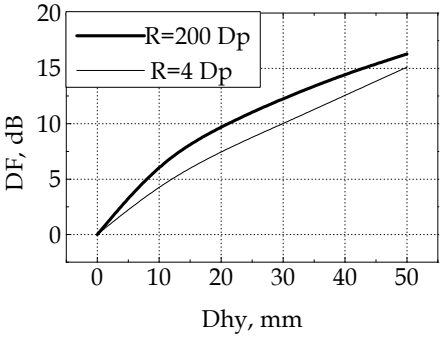


a) the deviation angle vs Dny

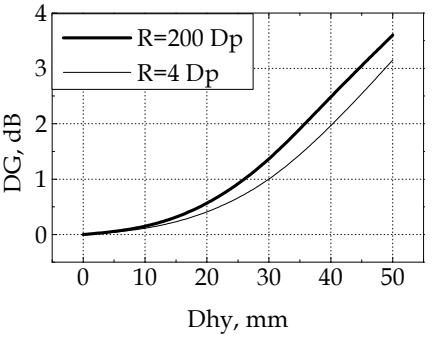


b) the coefficient of widening vs Dhy

Fig. 15. Scanning in a reflector antenna.



a) An increase of side lobe level at scanning



b) A decrease of gain at scanning

Fig. 16. Changes of antenna parameters at scanning.

From fig. 16 it follows that the effects of increasing the side lobe level and decreasing the antenna gain at scanning are reduced with decreasing the distance to the focusing point.
With increasing D_{hy} the cubic phase error on the paraboloid aperture increases. This results in rapid growth of side lobe levels. Therefore the scanning sector is not great large. It is necessary to decrease a cubic phase error for widening a scanning sector. It is can be made possible by shifting feed-horn along the focal axis additionally by D_{hz0} values. It is demonstrated by the example of the multi-beam antenna with parameters: D_p =300mm; F_p =150mm at frequency 37 GHz. This antenna is focused on the distance $R = 4D_p$ =1200mm. Every beam conforms to one feed-horn in the antenna. The number of feed-horns is $N_h = 30$. The feed-horns are located symmetrically in relation to the focal axis. The coordinates of feed-horns aperture centers along Y-axis are marked as D_{hyn}, on Z-axis – D_{hzn} are. The feed-horns numbers change from -15 up to 15, $-15 \le n \le 15$. The horns with the coordinate D_{hyn} <0 have numbers $-15 \le n < 0$, the feed-horns with coordinates Dhym>0 have numbers $0 < n \le 15$. The feed-horns location along Y-, Z-axis and their size are optimized to the criteria of minimum of side lobe levels and criteria of nearbouring antenna diagram crossing at -3dB level. The antenna diagrams conforming to horns $-1 \le n \le 15$ are shown in figure. 16. The nearest to focal axis horns have numbers $n = \pm 1$.

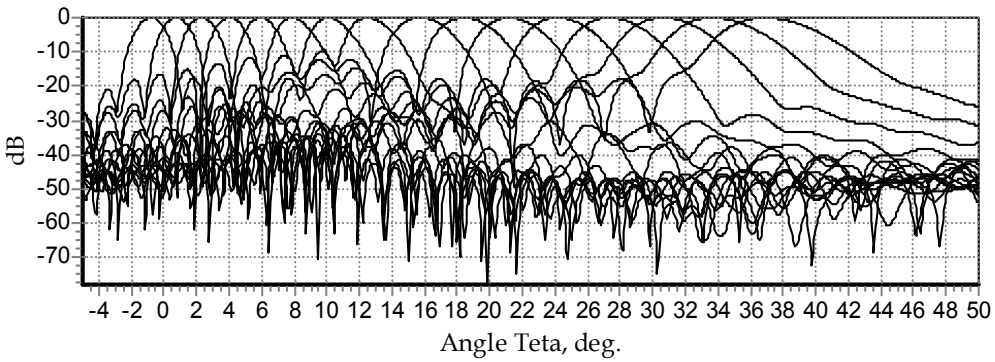


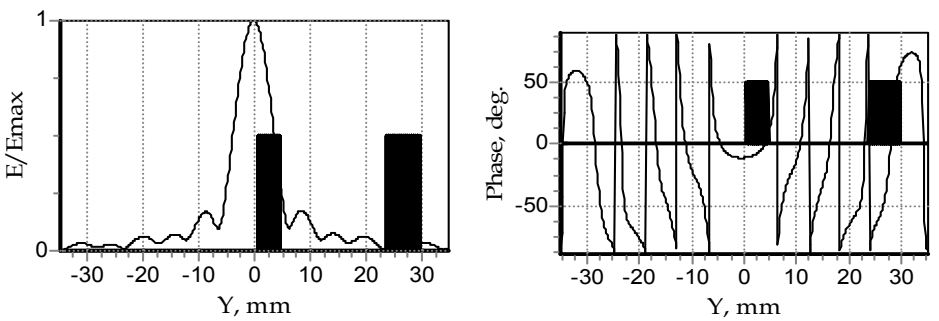
Fig. 17. The antenna diagrams of multi-beam antenna.

From fig. 17 it follows, that by optimization of the feed-horns sizes and location the sector taken up by beams can be essentially widened in comparison with the sector of scanning.
Multi-beam reflector antennas are used in radioimaging systems functioning in the passive mode. In this mode an antenna receives a signals radiated by some object in the near–field or intermediate-field antenna zone. Every reception channel is formed one horn of the feed-horn and receives a signal from the element of allowance on the object. The main lobes of the antenna diagram of the neighboring channels cross at non-zero level (generally it's 0.007 to the maximum). Therefore a signal received from this by the feed-horn of this channel is overlapped by adjacent bin signals (i.e. the desired signal is overlapped by interference and that leads to image degradation). The level of quality degradation can be evaluated as the ratio of the power received from necessary bin on the object to the power received from the adjacent bin in this channel.
Feed-horn isolation depends on the distance between the aperture centers l; on the dimensions of the feed-horn A_h, B_h, R_h; the dimensions of the paraboloid D_p, F_p on;

frequency f. Further the main principles will be considered an with example of an antenna with D_p =300mm, F_p =150mm and f=37GHz. Antenna diagrams of different channels are shown in the fig.17. The example of an amplitude distribution and the field phase distribution in the focus plane Z=const for an antenna focused at the distance of R=4 D_p are shown in fig. 17. The spherical wave source point has the following coordinates X=0, Y=0, Z=R. Two feed-horns (depicted as two black triangles) with numbers n=1 and n=5 are also shown. The dimensions of these feed-horns apertures and the coordinates of their centers after optimization of antenna diagrams are the following:

Horn 1- A_h =10,5 mm; B_h =4 mm; D_{hyn} =2,5 mm; D_{hzn} =22 mm

Horn 2 – A_h =10,5 mm; B_h =6 mm; D_{hyn} =27,75 mm; D_{hzn} =26 mm



a) Distribution of amplitude b) Distribution of phase

Fig. 18. the distribution of the field amplitude and field phase in the focusing plane Z=const.

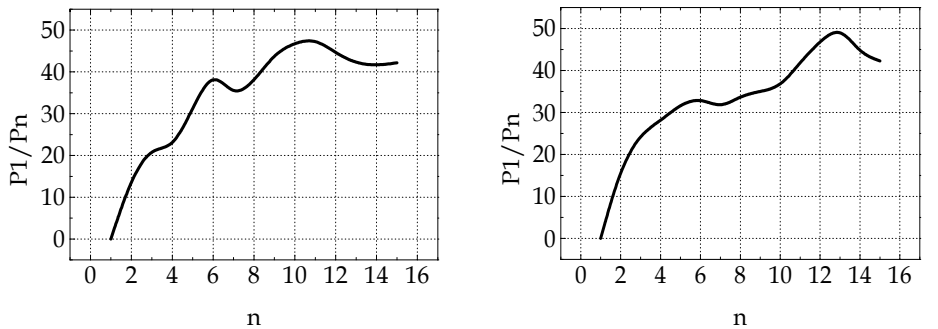
From fig.18 it follow that in the focusing plane the amplitude distribution and the field phase distribution are irregular, therefore as the distance between feed-horns increases the excitation amplitude of the feed-horn 2 doesn't decrease monotonous. Therefore the feed-horns isolation in a multi-beam changes monotonous with increasing the distance between feed-horns. The fig.18 depicts the dependence of the isolation coefficient between feed-horn 1 and feed-horns 2, 3,..., 15 depending on the feed-horn number in a multi-beam antenna. The antenna diagram of this antenna is shown in fig.16. The isolation coefficient is given by:

$$P1n = P1 / Pn \,, \tag{27}$$

Where P1 is the power received by the feed-horn1; Pn is the power received by the n-th feed-horn. The antenna receives the wave from the point on the Z-axis distant from the apex of the paraboloid at $R = 4D_p = 1200$ mm.

In fig.19 the similar dependence for an antenna with the feed-horn aperture dimensions and the position in the Y and Z-directions aren't modified according to a criterion of side lobe levels minimum; all the feed-horns have the same dimensions (A_h =10,5 mm; B_h =4 mm) and they are located in the focusing plane Z= D_{hzn} = 22mm. The feed-horns are located along the Y-axis equidistantly. The distance between the centers of the adjacent feed-horns equals 5 mm. At such distance the main lobes of feed-horns 1 and 2 cross each other at the level -3dB. The antenna diagrams conforming to the feed-horns numbered as n=-1…15 are shown in fig. 20.

From fig.16, 19 and 20 it follows that when the antenna receives a spherical wave from a point on the focal plane, the isolation coefficient between the feed-horn in the focusing point (the feed-horn 1) and the adjacent (feed-horn 2) is not less than 16-17 dB. The isolation coefficient at the feed-horn 1 with the rest feed-horns is not less than 20dB. If the feed-horns location in the focusing plane provides crossing the antenna diagrams at the level of -3dB then the level of the side lobes in antenna with the dimensions of the feed-horn corrected in the focusing plane and along the focal axis is distinctly lower than in an antenna without the dimension correction and positions of the feed-horns. The isolation of the feed horns in these two cases doesn't differ much.



a) with dimension correction b) without dimension correction

Fig. 19. The dependence of isolation coefficient on feed-horn number.

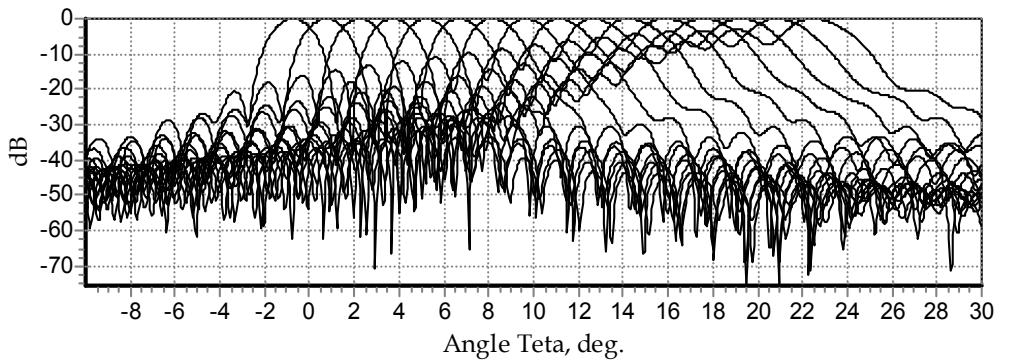


Fig. 19. The antenna diagram of multi-beam antenna without correction of feed-horns dimensions and its location.

When the antenna receives a spherical wave from the point located at $Y \neq 0$ some peculiarities occur. It is connected with the fact that the distribution of field amplitude in the focusing plane is asymmetrical regarding the maximum. Therefore the isolation coefficient between the feed-horn concerned and adjacent feed-horn located to the left and to the right from it is different.

Monopulse double-reflector Cassegrain antenna.

Monopulse antennas are designed for forming two differential diagrams and one sum diagram. The schemes for construction of monopulse double-reflector antennas (MDA) are described in literature. The simplified scheme of a Cassegrain antenna is shown in fig. 22. Designations: 1- is paraboloid (reflector); 2 is a hyperboloid (sub-reflector); 3 – is monopulse horn. The antenna receives a spherical electromagnetic wave coming from the point P (R,0,f) of the far-field zone with angular coordinates 0, f. On the fig. 22 one beam in the receiving mode is depicted in fig. 22 with dotted lines.

The feed-horn must generate three signals: two differential and one summary. Further the numerical modeling results are given and the main principles are described for two types of MDA: a) with a feed-horn in the form of 4-horns executed by TE01 wave mode; b) with a multimode feed-horn. The numerical modeling is produced for antennas with paraboloid diameters the D_p =30λ and the ratio F_p / D_p =0,4. The diameter of the hyperboloid depends on the eccentricity "e" and on the far focal distance Fg1.

A multimode feed-horn is a pyramidal horns with waves TE_{10}, TE_{20}, TE_{11}, TM_{11}. The scheme of simultaneous excitation of these wave modes though the isolated inputs is described in literature. These waves form necessary antenna diagrams: the wave TE_{10} form summary antenna diagram, the wave TE_{20} forms a differential antenna diagram in the magnetic plane (H plane); the sum of waves TE_{11} + TM_{11} forms the differential diagrams in the electrical plane (E plane). The parameters of the antenna diagram of the sum and differential channels depend on the dimensions of the feed-horn cross-section A_h, B_h and the eccentricity E.

Further the dependencies of antenna parameters from A_h, B_h are analyzed. The numerical modeling is produced for antennas with diameters of the paraboloid D_p =30λ and the ratio F_p / D_p =0,4.

The diameter of the hyperboloid depends on the eccentricity E and the distance between paraboloid apex and the feed-horn aperture. The hyperboloid must be inscribed in the aperture angle of paraboloid, as it is shown in fig. 6. The diameter of the hyperboloid D_g must not be larger than $0.25\,D_p$ to reduce the shadow effect. Further results of modeling are given for the case of H=3λ and D_g =0,22 D_P (the eccentricity e=2). For these parameters the level of side lobes in the sum channel in E and H planes is equal and it doesn't exceed -25dB; if the condition A_h =3,33λ; B_h =2,33λ (A_h / B_h =1,428) is fulfilled. The antenna diagrams of the sum channel for the mentioned parameters are shown in fig. 22; in the plane E with a thick line and for H plane with a thin line. The A_h, B_h values, the diameters of hyperboloid D_g and antenna parameters in the summary and differential channels depend on the hyperboloid eccentricity.

The side lobes level and the width of the main lobe of antenna diagram for the summary and differential channels depend on A_h/λ and B_h/λ ratios. With growth of the eccentricity the hyperboloid diameter decreases if the condition of equality of the paraboloid aperture angles and the hyperboloid aperture angle from the focus of paraboloid (the near focus of hyperboloid) is fulfilled. The dependence of the ratio D_g / D_p on the eccentricity E with A_h =3,33λ and B_h =2,33λ is shown in fig. 23. This figure also depicts the dependence of the antenna efficiency factor (Kef) in the summary channel from the hyperboloid eccentricity. The antenna efficiency factor is the value that connects the antenna gain (G), the

area of paraboloid aperture (S) and the wavelength. This relation is determined from well-known formula:

$$G = 4\pi \cdot S \cdot Kef / \lambda^2, \text{ the } S = \pi D_p^2 / 4. \quad (28)$$

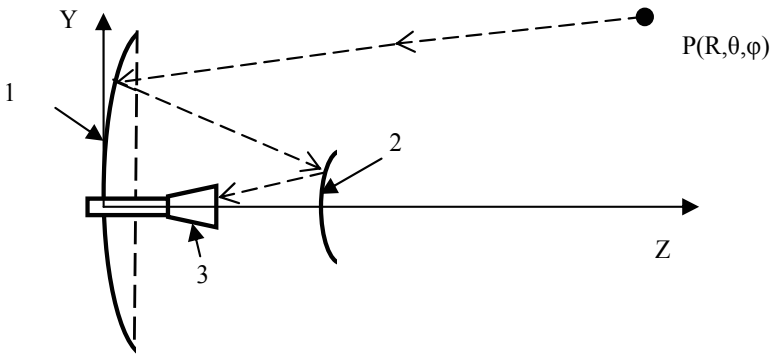


Fig. 22. An Cassegrain antenna scheme.

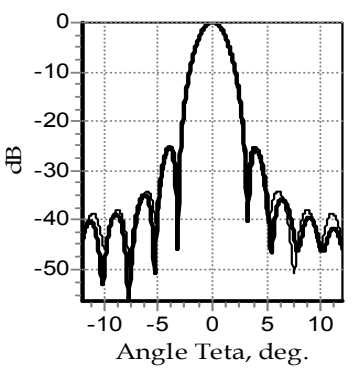


Fig. 23. The antenna diagram for the sum

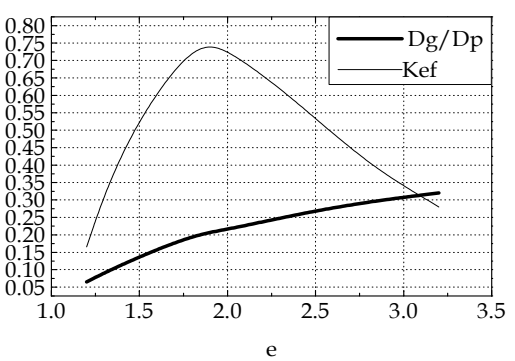
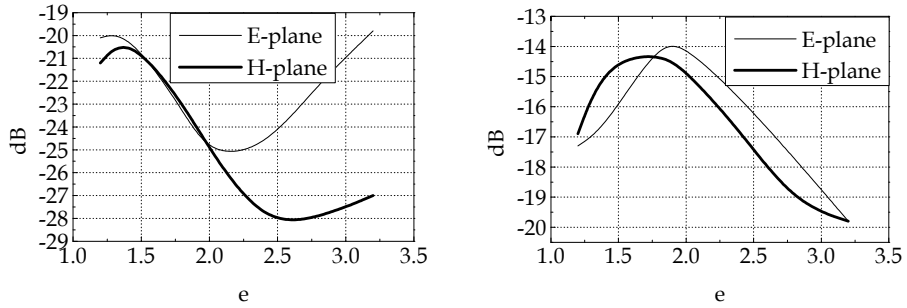


Fig. 24. The D_g / D_p and Kef versus eccentricity and differential channels.

Depending on the eccentricity the side lobe of summary and differential channels in the E- and H- planes for A_h =3,33λ, B_h =2,33λ change as it is shown in fig. 24. In our calculation we used a lens in the horn to achieve equal phase distribution in the aperture.
With the growth of the eccentricity the hyperboloid diameter increases. Therefore if the values A_h, B_h remain the same the horn field level decreases at the edge of hyperboloid surface. It leads to reducing the side lobes level and to increasing the main lobe width. At the expense of the side lobes reducing the antenna gain increases. At the expense of the main lobe extension the antenna gain decreases.
These two factors lead to the situation when at a certain level of the horn field at the edge of the hyprboloid the antenna gain reaches its maximum. In this case the antenna gain is maximal. Numerical analysis indicates that the maximum of antenna gain in the summary channel is observed when the field level at the edge of the hyperboloid in comparison with the center is $\Delta \approx 0{,}3$.



a) sum channel b) differential channels

Fig. 24. Side lobe levels versus hyperboloid eccentricity

The dependence of the efficiency factor and the antenna gain in differential channels on the eccentricity is shown on the fig. 24. The maximum of Kef is observed when the level on the edge of the aperture in the plane E is $\Delta_e \approx 0{,}4$; in the plane H is $\Delta_h \approx 0{,}3$.

From fig. 22 it follows that for reducing of side lobes level in the summary channel it's necessary to increase eccentricity in comparison with e=2 besides side lobes in the summary channel in the E- and H-planes are equal. The efficiency factor increases, if E<2.5.
But then the width of main lobe in the summary channel is increased and the antenna efficiency coefficient for the summary channel is decreased.

4. Conclusion

In the chapter the mathematical model of a reflector antenna is described with using the physical optical method and the waveguide excitation theory. New results of research of regularity and parameters of reflector antennas are:

- The field distribution in a near-field zone in a tangent plane and along a focal axis in radiation mode and receiving mode at different wave dimensions of antenna elements; in the receiving mode the antenna is illuminated by a spherical wave from the point with known coordinates, located in any space zone;

- The isolation between channels in a multi-beam antenna in receiving mode of spherical wave, radiated from near-field zone point depending on geometrical parameters of antenna elements.
- Changing diagram parameters of a reflector antenna during scanning and focusing in to given distance.

The results obtained can be used with designing a different purpose antenna.

5. References

Charles, M. An Extension of Rushh's Asymptotic Physical Optics Diffraction Theory of a Paraboloid Antenna. IEEE Trans. Antennas Propagat., vol.23, pp.741-743, Sept., 1975.

Chen, J & Xu, Y. Analysis and Calculation of Radiation Patterns of Cassegrain Antennas. IEEE Trans. Antennas Propagat., vol. 38, pp. 823 – 830.

Fitzgerald, W. The Efficiency of Near-Field Cassegrainian Antennas. IEEE Trans. Antennas Propagat., AP-14, no.9, pp.648-650, Sept. 1972.

Hannan, P.W. *Optimum feeds for all three modes of a monopulse antenna I: Theory*, IRE Trans. Antennas Propagat., vol. 9, pp. 444 - 454, September 1961.

Houshmand, D; Lee, S-W; Rahmat-Samii, Y & Lam, P. Analysis of Near-Field Cassegrain Refletctor: Plane Wave Versus Element-by-Element Approach. 1988 IEEE Int. Antennas Propagat. Symp. Dig. vol. 26, pp. 124 - 127, June 1988.

Khayatian, B; Rahmat-Samii, Y; *Characteristics of dual reflector antennas with gaps placed on the subreflector: MoM and PO analysis*, 1999 IEEE Int. Antennas Propagat. Symp. Dig. vol. 37, pp. 2332 - 2335, June 1999.

Laybros, S; Combes, P.F.; &Mametsa, H.J. The «Very-Near-Field» Region of Equiphase Radiating Apertures. IEEE Antennas & Propagation Magazine. 2005, vol. 47, No.4, pp.50-66.

Narasimhan,M; Ramanujam,P; & Raghavan, K. *GTD analysis of a hyperboloidal subreflector with conical flange attachment*, IEEE Trans. Antennas Propagat., vol. 29, pp. 865 - 871, November 1981.

Narasimhan, N and Christopher S. A New Method of Analysis of the Near and Far Fields of Paraboloidal Reflectors. IEEE Trans. Antennas Propagat., AP-32, no.1, pp.13-19, January 1984.

Narasimhan M & Govind, K. Front-to-back ratio of paraboloidal reflectors, IEEE Trans. Antennas Propagat., vol. 39, pp. 877 - 882, July 1991.

Rahmat-Samii, Y. *Subreflector extension for improved efficiencies in Cassegrain antennas--GTD/PO analysis*, IEEE Trans. Antennas Propagat., vol. 34, pp. 1266 - 1269, October 1986.

Rahmat-Samii, Y; *Jacobi-Bessel analysis of reflector antennas with elliptical apertures*, IEEE Trans. Antennas Propagat., vol. 35, pp. 1070 - 1074, September 1987.

Radar Handbook. Editor-In-Chif Skolnik M.I. McGraw_Hill Book Company, 1970.

Rusch, W. Physical-optics diffraction coefficients for a paraboloid. Electronic Lett., vol.10, pp.358-360, Aug.22, 1974.

Valentino, A & Toulios,P. Fields in the Focal Ragion of Offset Parabolic Antennas. IEEE Trans. Antennas Propagat., vol. 24, pp. 859 - 865, November 1976.

Watson, W. The Field Distribution in Focal Plane of a Paraboloidal Reflector. IEEE Trans. Antennas Propagat., vol. 12, pp. 561 - 569, September 1964.

10

Modeling of Microwave Heating and Oil Filtration in Stratum

Serge Sysoev[1] and Anatoli Kislitsyn[2]
[1]*Surgut State University,*
[2]*Tyumen State University*
Russia

1. Introduction

Extraction of high-viscosity oil and bitumen is an important practical problem, because the reserves of such deposits are significant, and their role in common stocks of organic raw materials is constantly increasing. However, on account of the high viscosity of oil, and but also because of frequent blockage bottomhole zone due to sediments of colloidal surface-active components of oil extraction is becomes possible only after preliminary heat treatment of the stratum. Traditional methods of thermal treatment - hot steam or hot liquid - are in this case ineffective. Moreover, their widespread use may lead to severe environmental consequences in the form of violations of the hydrogeological environment.

One of the promising methods of thermal treatment is an electromagnetic heating of the productive layers. Due to deep penetration and the volumetric heat release, and absence of coolant, electromagnetic radiation can provide (compared to traditional methods) high speed and uniform heating, the possibility of optimal control and automation of technological processes, virtually eliminate the harmful effects on the environment. The results of laboratory and field trials in Russia (Sayakhov et al., 1970, Makogon et al., 1989, Sayakhov et al., 2002) and practical experience of using this technology on an industrial scale in the U.S. and Canada (Da Mata et al., 1997, Vermeulen & McGee, 2000, Sahni et al., 2000, Chhetri & Islam, 2008) show perspective utility of this trend. However, the effective realization of these opportunities is hindered by the lack of reliable data on the study of heat and mass transfer processes in multiphase media, typical for the oil and gas technologies, when subjected to these media microwave electromagnetic radiation. The main objective is to determine optimal modes of stimulation, namely: the frequency and power of a source of microwave radiation, the parameters of the antenna, the possibility of using nonlinear properties of the medium to enhance impact on the models as close as possible to real conditions.

In Russia, work on the effects of high frequency electromagnetic radiation on the oil reservoir was started in the late 60-ies by a team from Bashkir State University under the leadership of F.L. Sayakhov (Sayakhov et al., 1970, Sayakhov et al., 1975). Their industrial-scale plant was successful tested at Sushuglinsk' and Mordovo-Karmalsk' oilfields. At this industrial-scale plant, the electromagnetic energy from the high-frequency generator for the feeder (two coaxial tubes) is inserted into the well. The outer sheath of coaxial cable joins the casing and the central thread cable - to the tubing at a depth of about 5 meters, so that the

upper part of the column and tubing formed the short-circuited line, equal to 1/4 wavelength. Inside the tubing is submerged rod pump. Casing and tubing are used as a coaxial transmission line for supplying a high frequency electromagnetic energy to the radiator. The radiator consists of the lower casing and the bottom of the tubing, which stands below the casing with 1/4 wavelength (quarter-wave linear radiator). 2.5-inch duralumin tubes were used as tubing. A diameter of casing is 9 inches. Insulating washers (plastic ring 15 mm thick) placed every 8-10 m along the tubing, were used for insulation of tubing from the casing. Generator with an operating frequency of 13.56 MHz and output power (under optimal conditions) 63 kW was used as a source of high frequency electromagnetic radiation. The average yield before the start of heating was 0.1 m3/day; water cut was 30%. As a result of heating at the output power of 20-30 kW steady-state temperature set at 110 C after 7 days; yield increased to 0.25 m3/day, i.e. 2,5 times; water cut was reduced to 7-8%, i.e. more than 3 times. A well operated with a high yield flow rate for 17 days after the end of the electromagnetic effects.

In the U.S.A., research of electromagnetic effects on the oil wells was started in the late 70's. These studies culminated in a series of successful tests on the oil fields of the United States and Canada, and current technology of high-frequency electromagnetic radiation is reduced to a cost-effective and competitive level that allowed to move to its practical use in industrial scale. The effect of electromagnetic heating of the near-well zone can be illustrated by the tests carried out on the field in Alberta. To assess the effectiveness of the heating for this field computer simulations were carried out. The modeling predicted approximately twofold increase of well production by electromagnetic effects. The well was drilled in January 1986, and in March 1986 started production of oil. Up to the impact well gave about 6 barrels of oil per day. A month after the start of operation was launched electromagnetic heating. A few days later oil production increased and set at about 20 barrels per day, i.e. even higher than had been predicted by numerical simulation.

Successful tests were conducted in several other fields of the United States and Canada (in Oklahoma, Utah, Texas, California). Currently in the U.S.A. (New Jersey) the company Global Resource Corp. has been successfully working in this area. This company is a developer of a patented microwave technology and machinery that extracts oil and petroleum products from shale deposits, tar sands, capped oil wells, coals and processed materials such as tires and plastics as well as dredged soil from harbors and river bottoms. This process produces significantly greater yields and lower costs than are available using existing technologies.

Over the past 10-15 years several reviews of methods of electromagnetic heating for enhanced oil recovery (Da Mata et al., 1997, Vermeulen & McGee, 2000, Sahni et al., 2000, Chhetri & Islam, 2008) have been published.

Theoretical studies of heat and mass transfer in the oil stratum under the influence of high frequency electromagnetic radiation were carried out by teams of specialists under the leadership of R.I. Nigmatulin and F.P. Sayakhov (Zyunk Ngok Khai et al., 1987, Kislitsyn & Nigmatulin, 1990, Sayakhov et al., 1998, Sayakhov et al., 2002, Kovaleva et al., 2004).

The process of heating and filtration of bitumen in porous medium volume heat source, arising due to absorption of energy of electromagnetic waves was studied (Zyunk Ngok Khai et al., 1987). One-dimensional problem was solved taking into account the phase transition. They found a stationary solution for spherically symmetric source of electromagnetic waves and self-similar solution for a cylindrically symmetric source. Numerical simulation of the space heating and filtration of oil in the presence of a moving

front of melting was performed in the one-dimensional model. Quantitative estimates of the size of the heat zone were obtained. It was also pointed out to the danger that the source of too much power will cause overheating near wellbore zone. The negative consequences of such overheating are the decomposition of oil near the well, the deformation of the skeleton of porous rock, the destruction of wells, etc.

The theoretical study of heat and mass transfer in the oil stratum when it is heated by high-frequency electromagnetic radiation was performed on one- and two-dimensional models in research (Sayakhov et al., 1998, Sayakhov et al., 2002, Kovaleva et al., 2004, Kislitsyn, 1993, Kislitsyn, 1996). In these studies, considerable attention was paid to the propagation of electromagnetic waves in the oil reservoir and the distribution of the density of volumetric volume heat sources. The valuation of efficiency and cost effectiveness of the method in terms of energy balance has been made.

The filtration processes in porous media filled with a solid gas hydrate or liquid, with depression and thermal effects (including the electromagnetic heating), which leads to phase transitions (gas hydrate decomposition, boiling liquid) were studied in research (Shagapov & Syrtlanov, 1994). Optimal regimes of the heating stratum by using high-frequency electromagnetic radiation were determined to gas hydration control in the near wellbore zone.

Summarizing this brief review, it should be noted that there are a number of studies which examined the processes of heating and filtration of oil in stratums when exposed to high-frequency electromagnetic field. In these studies important results that may be used to estimate the depth and duration of heating and to select the optimal modes of exposure were obtained. Overall, however, the problem can not be well studied. In all the works cited above the equation for the electromagnetic field with the type of radiating antenna, temperature and frequency dependences of dielectric loss tangent of the medium wasn't used. Neglect of these circumstances can cause significant inaccuracies, and even erroneous results.

In this paper we propose a mathematical model closer to the actual conditions that includes two-dimensional system of interrelated equations of heat transfer and piezoconductivity, supplemented by the equation for the electromagnetic field with the type of radiating antenna, temperature and frequency dependence of dielectric loss tangent of the medium.

2. Model and equations

Numerical studies were performed with a two-dimensional axisymmetric model, a diagram of which is shown in Fig.1. The petroleum stratum 2 is contained between planes perpendicular to the z-axis (1 - cap rock, 3 - underlying bedrock). The plate is bounded above and below by an infinite medium, the physical characteristics of which (thermal conductivity, density, heat capacity) differ from those of the plate. An electromagnetic radiation source 4 with an antenna is placed in the well. In this model, the antenna consists of a coaxial cable with a ring-shaped slot 7 cut on the outer conductor 6 from the short-circuited tip (5 - the central conductor of coaxial cable). The isolines of magnetic strength 8 in the coaxial cable, the petroleum stratum and the adjacent rock are shown in Fig.1.

Electromagnetic waves propagate in a radial direction about the well; they are absorbed and volume heating of the plate and adjacent rock occurs. Because of the heating the viscosity of the oil decreases and its flow into the well increases.

For fixed source power the size of the heated zone depends on the physical parameters of the medium and the electromagnetic wave penetration depth. This depth, in turn, depends

on the frequency of the radiation and can thus be controlled. For too great penetration depths (too low a frequency) the source energy is dissipated in a large region and leaks into the adjacent rock without producing the required heating. For too small a penetration depth (too high a frequency) intense heating of a small region surrounding the source occurs, a high temperature gradient develops, and heat is lost intensely upward and downward without providing the required radial heating. In both cases the heated zone is small and heating is ineffective. Consequently, there must exist some optimum frequency at which (for fixed source power) the most effective heating can be produced. As for source power, within the framework of the model used, the higher that power, the higher the well yield, but also the higher the heat loss. Therefore the efficiency of heating (ratio of the increase in petroleum yield to energy expended) can prove low for too high power level. Moreover, the radiated power is limited by the fact that it is undesirable to heat the oil above the temperature at which it decomposes. Determination of optimum values for radiation frequency and power is the basic task of our numerical modeling.

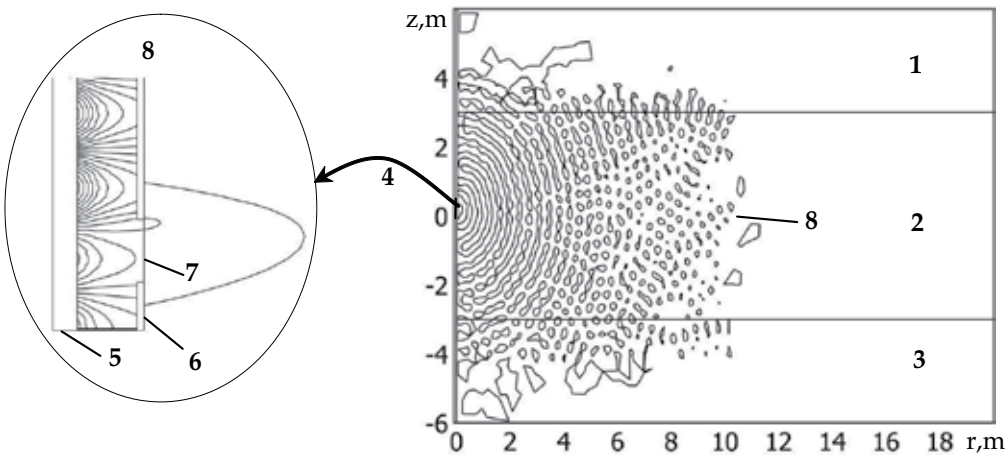


Fig. 1. A diagram of the model: 1 - cap rock; 2 - petroleum stratum; 3 - underlying bedrock; 4 - an electromagnetic radiation source with an antenna; 5 and 6 - a central conductor and an outer conductor of coaxial cable, respectively; 7 - a ring-shaped slot; 8 - isolines of magnetic strength

The model takes advantage of the problem's rotational symmetry, which allows modeling in 2D using cylindrical coordinates as indicated in Fig.1. When modeling in 2D, we can select a fine mesh and achieve excellent accuracy. The model uses a frequency-domain problem formulation with the complex-valued azimuthal component of the magnetic field as the unknown.

The radial and axial extent of the computational domain is in reality larger than indicated in Fig.1. This problem does not model the interior of the metallic conductors, and it models metallic parts using boundary conditions, setting the tangential component of the electric field to zero.

An electromagnetic wave propagating in a coaxial cable is characterized by transverse electromagnetic fields (TEM). Assuming time-harmonic fields with complex amplitudes containing the phase information, the appropriate equations are

$$\vec{E} = \vec{e_r}\frac{C}{r}e^{j(\omega t - k_0 z)}, \tag{1}$$

$$\vec{H} = \vec{e_\varphi}\frac{C}{rZ}e^{j(\omega t - k_0 z)}, \tag{2}$$

$$\vec{P}_{av} = \int_{r_{in}}^{r_{out}} \mathrm{Re}\left(\frac{1}{2}\vec{E}\times\vec{H}^*\right)2\pi r dr = \vec{e_z}\pi\frac{C^2}{Z}\ln\left(\frac{r_{out}}{r_{in}}\right), \tag{3}$$

where z is the direction of propagation, and r, φ, and z are cylindrical coordinates centered on the axis of the coaxial cable. $\vec{P}_{av}$ is the time-averaged power flow in the cable, Z is the wave impedance in the dielectric of the cable, while r_{in} and r_{out} are the dielectric's inner and outer radii, respectively. Further, ω denotes the angular frequency. The propagation constant, k_0, relates to the wavelength in the medium, Λ, as

$$k_0 = \frac{2\pi}{\Lambda} \tag{4}$$

In the stratum, the electric field also has a finite axial component whereas the magnetic field is purely in the azimuthal direction. Thus, we can model the antenna using an axisymmetric transverse magnetic (TM) formulation. The wave equation then becomes scalar in H_φ:

$$\left(\nabla\times\left(\left(\varepsilon_r - \frac{j\sigma}{\omega\varepsilon_0}\right)^{-1}\nabla\times\vec{H}\right)\right)_\varphi - \mu_r k_0^2 H_\varphi = 0, \tag{5}$$

where ε_r is the relative electric permittivity of the stratum; σ is the conductivity of the stratum; μ_r is the relative magnetic permittivity of the stratum.
The boundary conditions for the metallic surfaces are

$$\vec{n}\times\left(\vec{E_1} - \vec{E_2}\right) = 0, \tag{6}$$

where $\vec{n}$ is the normal to the surface, the inferior indexes 1 and 2 relate to the stratum and the adjacent rock, respectively.
The feed point is modeled using a port boundary condition with a power level set to several tens of kilowatts. This is essentially a first-order low-reflecting boundary condition with an input field $H_{\varphi 0}$:

$$\left|\vec{n}\times\sqrt{\varepsilon}\vec{E}\right| - \sqrt{\mu}H_\varphi = -2\sqrt{\mu}H_{\varphi 0}, \tag{7}$$

where

$$H_{\varphi 0} = \frac{\sqrt{\dfrac{\vec{P}_{av}Z}{\pi r\ln\left(\dfrac{r_{out}}{r_{in}}\right)}}}{r} \tag{8}$$

for an input power of $\vec{P}_{av}$ deduced from the time-average power flow.
The antenna radiates into the stratum where a damped wave propagates. As we can discretize only a finite region, we must truncate the geometry some distance from the antenna using a similar absorbing boundary condition without excitation. Apply this boundary condition to all exterior boundaries. Finally, apply a symmetry boundary condition for boundaries at $r = 0$:

$$E_r = 0, \ \frac{\partial E_z}{\partial r} = 0 \quad . \tag{9}$$

The volume heat source density is equal to the resistive heat generated by the electromagnetic field:

$$q(r,z,T,t) = \frac{1}{2}\mathrm{Re}\left[\left(\sigma - j\omega\varepsilon_r\right)\vec{E}\cdot\vec{E}^{*} \right], \tag{10}$$

where

$$\sigma = \varepsilon_0\omega\varepsilon'' = \varepsilon_0\omega\varepsilon_r tg\delta, \quad \varepsilon = \varepsilon_r - j\varepsilon'' \ , \tag{11}$$

where ε'' is the imaginary part of the relative electric permittivity, $tg\delta$ is the dielectric loss tangent.
The heat equation describes the nonstationary heat transfer problem:

$$c\rho\frac{\partial T}{\partial t} + \nabla\cdot\left(-\lambda\nabla T\right) + mc_1\rho_1\vec{v}\cdot\nabla T = q(r,z,T,t) \tag{12}$$

where T is the temperature of the medium; c , ρ , λ are the specific heat capacity, density and thermal conductivity of the medium, averaged over all phases (these quantities are different in the plate and adjacent rock, and are thus functions of z); c_1 , ρ_1 are the heat capacity and density of the filtering liquid (petroleum); m is the porosity coefficient; $\vec{v}$ is the filtration velocity vector.
The process of oil filtration is described by equation of piezoconductivity:

$$\frac{\partial p}{\partial t} = \frac{1}{m\beta_p}\nabla\cdot\left(\frac{k}{\eta}\nabla p\right) + \frac{\beta_T}{\beta_p}\frac{\partial T}{\partial t}, \tag{13}$$

where p is the pressure, k is the permeability coefficient, η is the viscosity of the filtering liquid (petroleum), β_p is the compressibility coefficient, β_T is the thermal expansion coefficient of the filtering liquid.
Equations (12) and (13) are interrelated in that Eq. (12) considers convective heat exchange, which is dependent on pressure (Darcy's law):

$$\vec{v} = -\frac{k}{\eta}\nabla p \ , \tag{14}$$

while Eq. (13) considers the dependence of the oil viscosity on temperature and its volume expansion due to heating. Natural convection in the gravitational field cannot develop under the given conditions, since the Rayleigh number

$$Ra = g\frac{\beta_T \rho_1^2 c_1 k}{\mu\lambda}\Delta T H \ll 1 \tag{15}$$

for any reasonable temperature head (H is the plate height).
The process of paraffin melting is accounted for in the following manner. It is assumed that the heat capacity within the stratum exhibits a singularity at the phase transition temperature T_s:

$$c(T) = c_0 + L\delta(T - T_s) \tag{16}$$

(L is the latent heat of phase transition and δ represents the delta function, which in numerical calculations is replaced by a "step" of finite width $2\Delta T_s$). Since the heatcapacity values for temperatures below and above T_s are different (c_0 and c_1, respectively), we can write the function $c(T)$ in the form

$$c(T) = \begin{cases} c_0 & when\ T < T_s - \Delta T_s, \\ \dfrac{c_0 + c_1}{2} + \dfrac{L}{2\Delta T_s} & when\ T_s - \Delta T_s \le T \le T_s + \Delta T_s, \\ c_1 & when\ T > T_s + \Delta T_s. \end{cases} \tag{17}$$

Initial and boundary conditions are as follows:

$$\begin{aligned} &T\big|_{t=0} = T_0; \\ &\left.\frac{\partial T}{\partial r}\right|_{r=b} = 0, \quad \left.\frac{\partial T}{\partial r}\right|_{r\to\infty} \to 0, \quad \left.\frac{\partial T}{\partial z}\right|_{z\to\pm\infty} \to 0; \\ &p\big|_{t=0} = p_0,\ p\big|_{r=b} = p_b,\ p\big|_{r\to\infty} \to p_0,\ \left.\frac{\partial p}{\partial z}\right|_{z=\pm H/2} = 0 \end{aligned} \tag{18}$$

(T_0, p_0 are the initial intraplate temperature and pressure, p_b is the pressure in the well, its radius is b).
Thus, the model is included in a system of two-dimensional interconnected equations of an electromagnetic wave propagating (5), heat transfer (12) and piezoconductivity (13) with appropriate boundary and initial conditions (18). The model takes into account phase transitions (process of paraffin melting) and temperature dependence of the dielectric loss tangent of oil.
In research (Kislitsyn & Fadeev, 1994) electric permittivity and dielectric loss tangent of certain types of high-viscosity oils (including Russian oil) in a wide range of frequencies and temperatures have been experimentally obtained. As a result, it was found that the dependence of complex permittivity ε on the frequency ω for oil is described by the model Havriliak-Negami (Havriliak & Negami, 1968):

$$\varepsilon = \varepsilon_r - j\varepsilon'' = \varepsilon_\infty + \frac{\varepsilon_s - \varepsilon_\infty}{\left[1 + (j\omega\tau_0)^{1-\beta}\right]^\gamma} + \frac{\sigma}{j\omega\varepsilon_0}, \quad 0 \le \beta < 1; \; 0 < \gamma \le 1, \tag{19}$$

where $\varepsilon_s, \varepsilon_\infty$ are static and high frequency limits of dielectric permittivity; τ_0 is the most probable relaxation time of molecules of the dielectric; β, γ are parameters characterizing respectively the width and asymmetry of the spectrum of relaxation times of the molecules of the dielectric. For γ = 1, this model goes into Cole-Cole model, and if more and β = 0, then the Debye model.

Sharing in the expression for ε the real and imaginary parts, we find

$$\varepsilon_r = \varepsilon_\infty + r^{-\gamma/2}(\varepsilon_s - \varepsilon_\infty)\cdot\cos(\gamma\vartheta), \tag{20}$$

$$\varepsilon'' = r^{-\gamma/2}(\varepsilon_s - \varepsilon_\infty)\cdot\sin(\gamma\vartheta) + \sigma/\omega\varepsilon_0, \tag{21}$$

where

$$r = \left[1 + (\omega\tau_0)^{1-\beta}\sin(\beta\pi/2)\right]^2 + \left[(\omega\tau_0)^{1-\beta}\cos(\beta\pi/2)\right]^2, \tag{22}$$

$$\vartheta = arctg\left[\frac{(\omega\tau_0)^{1-\beta}\cos(\beta\pi/2)}{1 + (\omega\tau_0)^{1-\beta}\sin(\beta\pi/2)}\right]. \tag{23}$$

The temperature dependence of ε_r and ε_2 are determined by the temperature dependence of the parameters β, γ, τ_0 and σ model.

Developed in research (Kislitsyn & Fadeev, 1994) method of processing experimental data allowed us to determine with good accuracy the model parameters Havriliak-Negami for various high-viscosity oil Tyumen region. The values of parameters allow us to describe the behavior of oil in the electromagnetic field in a wide range of frequencies and temperatures, in particular the important characteristics as the dielectric loss tangent $tg\delta = \varepsilon''/\varepsilon_r$, which affects the distribution of volume heat sources.

Figure 2 shows the dependence of dielectric loss tangent $tg\delta$ on temperature for oil of Russian field for a range of frequencies from 500 MHz to 2.4 GHz. The figure shows that with increasing oil temperature from the initial (T_0 = 293 K) to values of 330-360 K in the entire frequency range of the radiation the loss tangent increases approximately 1.5-fold, and then with further increase of temperature there is a decline of approximately 10 times when reaching decomposition temperature of the oil (about 530-550 K). Thus, the dependence of loss tangent with temperature for oil is nonlinear ("resonance") character, which significantly affects the process of heating oil electromagnetic radiation and should be considered when modeling this process.

Data on the viscosity of the Russian oil deposit depending on the temperature are obtained in (Kislitsyn & Fadeev, 1994). This dependence is well approximated by a generalized formula Andrade, which was used for modeling:

$$\eta(T) = \eta_\infty \exp\left\{E_\eta / R(T - T_s)\right\}, \tag{24}$$

where η_∞ is high-temperature limit of viscosity; E_η is activation energy of viscosity; T_s is temperature of complete solidification; R is universal gas constant.

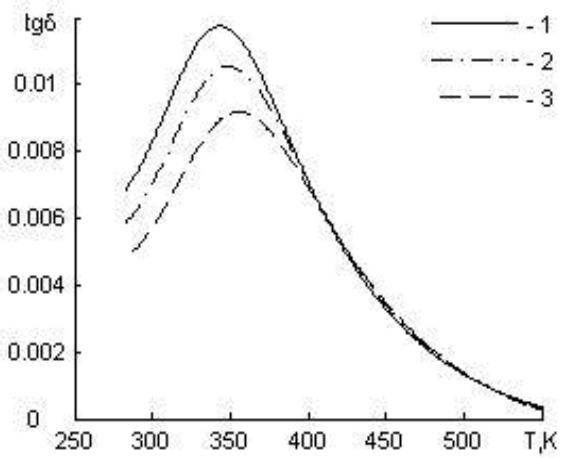


Fig. 2. The dependence of the dielectric loss tangent on temperature for oil Russian field for the frequencies: 1 - 500 MHz, 2 - 1 GHz, 3 - 2,4 GHz

Simulation of heating stratum was carried out by finite element commercial software package COMSOL Multiphysics. A numerical algorithm for the finite element method is based on the procedure of minimizing the functional corresponding to the continuous problem solved. The result of this procedure is the substitution of the system of partial differential equations system of algebraic equations with the coefficients approximating functions, which are actually the values of the unknown function at the vertices of the subdivision.

In the present research computational domain task was divided approximately into 40000 finite elements having the form of triangles. Finite element mesh was nonuniform. Concentration of elements was carried out in areas of expected strongest changes in temperature and electromagnetic field, i.e. near the radiation source and at the interfaces of the stratum-surrounding rock, where the size of finite elements was more than 10 times less than the wavelength of the radiation. As the basis functions piecewise-continuous quadratic Lagrange polynomials were used. The number of degrees of freedom of the problem was still approximately 170000. The numerical integration required to find the elements of the Jacobian, was carried out using the Gauss quadrature formula. To solve systems of linear algebraic equations was used Gaussian method, adapted to the use of very sparse matrices. The relative accuracy of calculations at each step of the iterative process was 0.01. Calculations were performed on a computer that has a processor with a clock speed of 3.33 GHz and 4 GB of RAM. Typical calculation time was approximately 60 hours.

In this research, a numerical study of electromagnetic heating oil stratum was carried out using physical parameters typical for heavy oil of the Russian Tyumen' field: oil density ρ_0 =940 kg/m^3, density of the rock stratum ρ_1 =2200 kg/m^3, density of the surrounding rocks ρ_2 =1580 kg/m^3, volume heat capacity of oil c_0 =2310 kJ/(m^3·K), average volumetric heat capacity of stratum c_1 =2310 kJ/(m^3·K), volumetric heat capacity of the surrounding

rocks c_2 =2310 kJ/(m^3·K), average thermal conductivity of the stratum λ_1 =1,0 W/(m·K), thermal conductivity of the surrounding rocks λ_2 =2,33 W/(m·K), average porosity of the reservoir 32%, melting heat L =160 kJ/kg. The values $\varepsilon_r(T)$, $\varepsilon''(T)$, $tg\delta(T)$, $\sigma(T)$ and $\eta(T)$, as a function of temperature, were determined by the above method.

2. Results and discussion

In this research the process of heating of stratum by electromagnetic radiation at frequencies f between 500 MHz and 2.4 GHz for 30 days was simulated. Heating time of stratum was chosen based on the fact that the typical heating time when using the traditional methods of heat treatment ranged from one to several months or even years. As a result of numerical study of the model based on equations (5), (12) and (13), supplemented by (10), spatial and temporal distribution of electromagnetic field, the volume density of electromagnetic energy, heat sources, temperature and viscosity were obtained. Some simulation results are shown in Fig. 3-6.

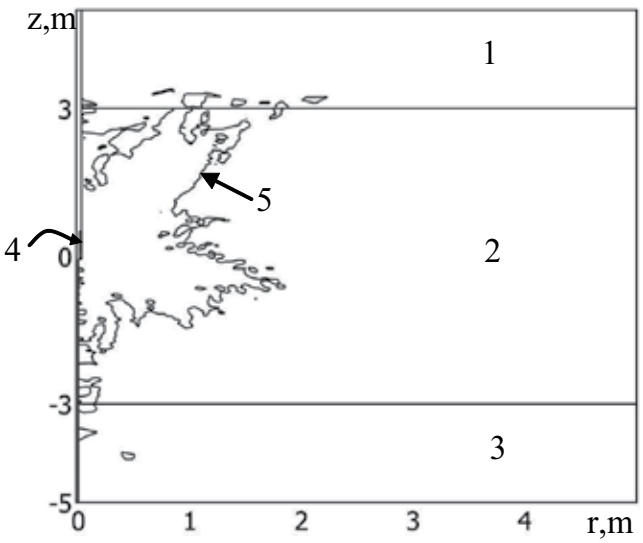


Fig. 3. The antenna diagram.
1 and 3 are top and bottom oil layer 2, 4 is a radiation source; 5 is isoline volume energy density of the electromagnetic field; radiation frequency f = 1 GHz and a power of W = 20 kW; time of heating the stratum is 10 days

The directivity and depth of penetration of radiation into the stratum can be judged by spatial distribution of volume energy density of the electromagnetic field, that is, in fact, this distribution characterizes the radiation pattern antenna. Figure 3 (1 and 3 are top and bottom oil layer 2, 4 is a radiation source) shows isoline volume energy density of the electromagnetic field 5 for the case of heating the stratum within 10 days of radiation source frequency f = 1 GHz and a power of W = 20 kW. Isoline 5 corresponds to the value of the energy density equal to $5 \cdot 10^{-6}$ J/m^3. The figure shows that the radiation pattern of antenna radiation, being axisymmetric, has a complex spatial distribution of electromagnetic field, its

form changes with time of heating stratum due to temperature changes in the electrical properties of the medium. By integrating the energy density over the respective volumes values of the energy of the electromagnetic field in the stratum and surrounding rocks were obtained. Comparison of these values showed that approximately 94% of this energy falls on the stratum and only 6% on the surrounding rocks, indicating that sufficient performance directional antenna is used.

The calculations of temperature fields in the oil stratum allowed to determine the maximum allowable power source at a given frequency of radiation and the heating time of stratum. Power of the radiation source is necessary to limit the value at which the maximum temperature of oil, corresponding to the beginning of its thermal decomposition (approximately 530-550 K) is reached. Figure 4 shows the results of calculations of temperature in the stratum, depending on the radial distance from the source, obtained within the proposed model in the cases without (curve 1) and with (curve 2) temperature dependence of loss tangent (the radiation frequency f = 1 GHz , source power W = 20 kW, heating time 30 days). Thus, accounting of the temperature dependence of loss tangent has a significant impact on the calculations of temperature fields near the source of radiation. This is due to the fact that with increasing of oil temperature above 420 K, the values of loss tangent are considerably smaller (10 times at T = 530 K) than its value at the initial temperature of the stratum (Fig. 2), and therefore, in accordance with expression (2), decreases in proportion to the density of volume sources of heat, which slows down the heating stratum. The results of the calculations showed that when the heating time of 30 days of stratum and the radiation frequency f = 500 MHz, the maximum permissible power source is W = 30 kW, when the radiation frequency f = 1 GHz - W= 20 kW, when the radiation frequency f = 2,4 GHz - W= 5 kW.

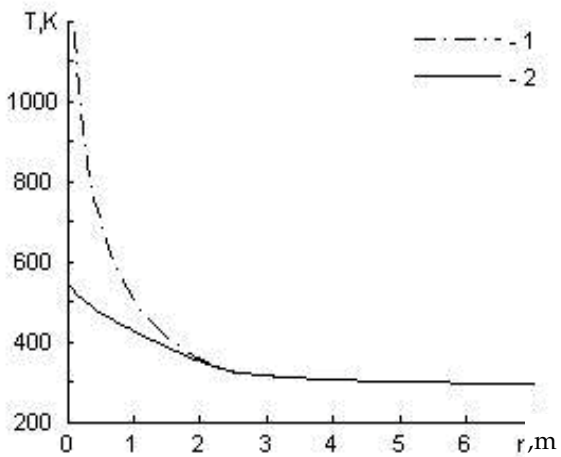


Fig. 4. The change of temperature in the stratum with the distance from the radiation source: without (curve 1) and with (curve 2) temperature dependence of loss tangent (the radiation frequency f = 1 GHz , source power W = 20 kW, heating time 30 days)

Figure 5 shows the isotherms of the temperature field after 10 days after the start of heating (source power W=20 kW, the radiation frequency f=1 GHz): curve 1 represents the isotherm

of 400 K, curves 2 and 3 are isotherms (323.05 K and 322.95 K), limiting the region of phase transition, 4 and 5 are isotherms 300 K and 294 K, respectively. In contrast to [7-9], where the phase transition is an infinitely thin front of melting, obtained in this study results indicate that under certain conditions, the extended region of phase transition (the area between the isotherms 2 and 3 in Figure 5) is formed. The distance at which the melting front moves along the axis r, and thus an important parameter of the process of heating - the volume of the melting zone - at a fixed heating time depends on the radiation frequency and power source.

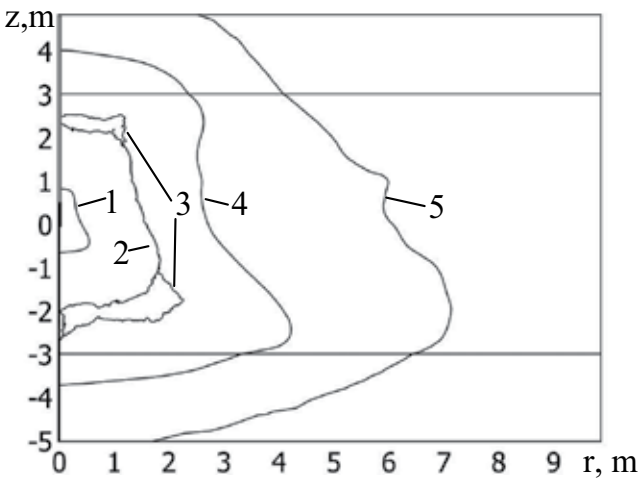


Fig. 5. The isotherms of the temperature field after 10 days after the start of heating (source power W=20 kW, the radiation frequency f=1 GHz): curve 1 represents the isotherm of 400 K, curves 2 and 3 are isotherms (323.05 K and 322.95 K), limiting the region of phase transition, 4 and 5 are isotherms 300 K and 294 K, respectively

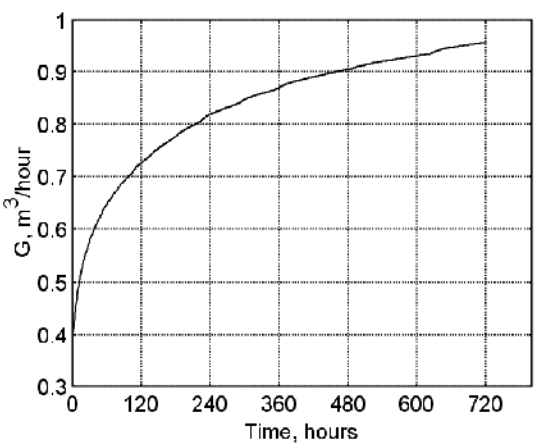


Fig. 6. Well yield as a function of heating time (frequency 1 GHz, power of source 30 kW)

Heating of oil reduces its viscosity, which, in turn, improves oil withdrawal. From a practical viewpoint the most important result of heating is the increase in well yield as compared to the yield of "cold" well. Figure 6 shows the dependence of increase in well yield as a function of heating time (optimal parameters for the Russian field: frequency 1 GHz, power of source 30 kW). The figure shows that this mode of heating leads to an increase well production by 2,3 times. At the same time energy costs account for about 60 kilowatt-hours per 1 m3 of additional oil production, which is quite acceptable from a practical point of view.

It has been shown that the efficiency of heating depends significantly on proper choice of radiator frequency and power. These results are quite usable from a practical standpoint, and the electromagnetic heating method is technically achievable and competitive, for example, with the in-situ combustion method. It is shown that high frequency microwave heating may be used for stimulating oil production high-viscous, low permeability stratum.

3. References

Sayakhov, F.L.; Babalyan, G.A. & Chistyakov, S.I. (1970). On the high-frequency heating bottomhole zone, *Neft. Khoz.,* No.12, p.49-66

Makogon, Yu.F.; Sayakhov, F.L. & Khabibullin, I.L. (1989). The method of extraction of non-traditional hydrocarbon, *Doklady Akademii Nauk, Vol.306,* No.4, p.941-943, 0869-5652

Sayakhov, F.L.; Kovaleva, L.A. & Nasyrov, N.M. (2002). Heat and mass transfer in the system well - the stratum under the electromagnetic effects of the massive oil deposits, *Inzhenerno-Fizicheskii Zhurnal,* Vol.75, No.1, p.95-99, 0021-0285

Da Mata,W.; Peuch, J.C. & Baudrand, H. (1997). An overview of the RF heating process in petroleum industry, *Proceedings of Microwave and Optoeiectronics Conference,* pp.28-33 Vol.1, ISBN 0-7803-4165-1, Brazil, Aug 1997, Natal

Vermeulen, F. & McGee, B. (2000). In situ electromagnetic heating for hydrocarbon recovery and environmental remediation, *J. Can. Pet. Technol.,* Vol.39, pp.25-29

Sahni, A.; Kumar, M. & Knapp, R.B. (2000). Electromagnetic heating methods for heavy oil reservoirs, SPE Paper 62550. *SPE/AAPG Western Regional Meeting,* Long Beach, CA, June 2000

Chhetri, A.B. & Islam, M.R. (2008). A Critical Review of Electromagnetic Heating for Enhanced Oil Recovery, *Petroleum Science and Technology,* Vol.26, pp. 1619-1631

Sayakhov, F.L.; Babalyan, G.A. & Almetev, A.I. (1975). A method for extraction of viscous oil and bitumen, *Neft. Khoz.,* No.12, pp.32-34

Zyunk Ngok Khai, Kutushev, A. G. & Nigmatulin, R. I. (1987). The theory of filtration for liquids in a porous medium with volumetric heating by means of high-frequency electromagnetic fields, *Prikl. Mat. Mekh.,* Vol.51, No.1, pp.29-38

Kislitsyn, A.A. & Nigmatulin, R. I. (1990). Numerical modeling of the process of heating a petroleum stratum by high frequency electromagnetic radiation, *Prikl. Mat. Mekh.,* Vol.31, No.4, pp.59-64

Sayakhov, F.L.; Kovaleva, L.A. & Nasyrov, N.M. (1998). Study of heat and mass transfer characteristics in the bottomhole zone in the case of injection of the solvent simultaneously with the electromagnetic effects, *Inzhenerno-Fizicheskii Zhurnal,* Vol.71, No.1, p.161-165, 0021-0285

Kovaleva, L.A.; Nasyrov, N.M. & Haidar, A.V. (2004). Mathematical modeling of high frequency electromagnetic heating of bottomhole zone of horizontal oil wells, *Inzhenerno-Fizicheskii Zhurnal,* Vol.77, No.6, p.105-111, 0021-0285

Kislitsyn, A.A. (1993). Numerical modeling of petroleum heating and filtration in a plate under the action of high-frequency electromagnetic radiation, *Prikl. Mat. Mekh.,* Vol.34, No.3, pp.97-103

Kislitsyn, A.A. (1996). Numerical modeling of high-frequency electromagnetic heating of a dielectric plug clogging a pipe, *Prikl. Mat. Mekh.,* Vol.37, No.3, pp.75-82

Shagapov, V.S. & Syrtlanov, V.R. (1994). Filtering boiling liquid in a porous medium, *Teplofizika vysokikh temperatur, Vol.32,* No.1, pp.87-93, 0040-3644

Kislitsyn,A.A. & Fadeev, A.M. (1994). Dielectric relaxation in heavy oil, *Zhurnal fizicheskoi khimii*, Vol.68, No.2, pp.340-343, 0044-4537

Havriliak, S. & Negami, S. (1968). A Complex Plane Analysis of α-Dispersions in Some Polymer Systems, *J. Polymer Sci.,* Part C, No.14, pp. 99-117

Part 3

Materials

11

Numerical Simulation of Elastic-Plastic Non-Conforming Contact

Sergiu Spinu, Gheorghe Frunza and Emanuel Diaconescu
Department of Applied Mechanics, University "Stefan cel Mare" of Suceava
Romania

1. Introduction

A fast algorithm for elastic-plastic non-conforming contact simulation is presented in this work. While the elastic response of a material subjected to load application is reversible, plasticity theory describes the irreversible behavior of the material in reaction to loading beyond the limit of elastic domain. Therefore, elastic-plastic response of contacting bodies to loading beyond yield strength is needed to assess the load-carrying capacity of the mechanical contact.

The modern approach in simulating elastic-plastic contact is based on the algorithm originally proposed by Mayeur, (Mayeur, 1996), employing Betti's reciprocal theorem. Although Mayeur developed a model for the three-dimensional problem, numerical implementation was restricted to two-dimensional case, due to lack of formulas for the influence coefficients.

Problem generalization is due to Jacq, (Jacq, 2001), and to Jacq et al. (Jacq et al., 2002), who advanced a complete semi-analytical formulation for the three-dimensional elastic-plastic contact. The algorithm was later refined by these authors, (Wang & Keer, 2005), who improved the convergence of residual and elastic loops. The main idea of their Fast Convergence Method (FCM) is to use the convergence values for the current loop as initial guess values for the next loop. This approach reduces the number of iterations if the loading increments are small.

Nélias, Boucly, and Brunet, (Nélias et al., 2006), further improved the convergence of the residual loop. They assessed plastic strain increment with the aid of a universal algorithm for integration of elastoplasticity constitutive equations, originally proposed by Fotiu and Nemat-Nasser, (Fotiu & Nemat-Nasser, 1996), as opposed to existing formulation, based on Prandtl-Reuss equations, (Jacq, 2001). As stated in (Nélias et al., 2006), this results in a decrease of one order of magnitude in the CPU time.

Influence of a tangential loading in elastic-plastic contact was investigated by Antaluca, (Antaluca, 2005). Kinematic hardening was added by Chen, Wang, Wang, Keer, and Cao, (Chen et al.,2008), who advanced a three-dimensional numerical model for simulating the repeated rolling or sliding contact of a rigid sphere over an elastic-plastic half-space.

The efficiency of existing elastic-plastic contact solvers, (Jacq et al., 2002; Wang & Keer, 2005) is impaired by two shortcomings. Firstly, the algorithms are based on several levels of iteration, with the innermost level having a slow convergence. Secondly, the effect of a three-dimensional distribution in a three-dimensional domain, namely residual stresses related to plastic strains, is computed using two-dimensional spectral algorithms.

A numerical approach to simulate the elastic-plastic contact, based on Betti's reciprocal theorem, is overviewed in this work. Computation of residual stresses due to plastic strains is accelerated by implementing three-dimensional spectral methods, in a hybrid convolution-correlation algorithm. Pressure-free surface condition in Chiu's inclusion problem decomposition is imposed with the aid of Boussinesq fundamental solutions and superposition principle. The newly proposed algorithm appears well adapted to numerical simulation of elastic-plastic contacts. Fotiu and Nemat-Nasser's universal algorithm is employed to derive plastic strain increment. The convergence of the residual part is therefore improved dramatically, and computationally intensive residual stress assessment is moved to an upper iterative level, allowing for finer resolutions in problem digitization.

2. Formulation of continuous elastic-plastic contact problem

Since the works of Mayeur, (Mayeur, 1996), and Jacq, (Jacq, 2001), Betti's reciprocal theorem is used in elastic-plastic contact modeling to assess surface normal displacement and stress state in an elastic half-space in the presence of plastic strains. The basis of Betti's theorem is the equality between the work done by the virtual force through the displacements produced by the real force and the work done by the real force through the displacements produced by the virtual force.

According to this formulation, if two independent loads are applied to an elastic body of volume Ω and of boundary Γ, generating two independent states (u,ε,σ) and $(u^*,\varepsilon^*,\sigma^*)$ with vanishing body forces, and the latter corresponds to a unit load applied along the direction of $\vec{x}_3$, in a point A of the boundary (a unit impulse):

$$p_3^*(M)=\begin{cases} 0, & M\neq A; \\ \left(dx_1dx_2\right)^{-1}, & M=A, \end{cases} \tag{1}$$

the following equation holds:

$$u_3(A)=\int_{\Gamma_C} u_{33}^*(M,p_3^*(A))p_3(M)d\Gamma+2\mu\int_{\Omega_p}\varepsilon_{ij}^p(M)\varepsilon_{3ij}^*(M,p_3^*(A))d\Omega. \tag{2}$$

Here, Γ_C is the boundary subdomain with normal tractions p_3 defined, and Ω_p the volume subdomain with existing plastic strains ε^p, both corresponding to state (u,ε,σ), μ Lamé's constant and M the integration point. This point is located within Γ_C in the first term of Eq. (2) and within Ω_p in the second. Consequently, $u_{33}^*(M,p_3^*(A))$ is the displacement in the direction of $\vec{x}_3$, and $\varepsilon_{3ij}^*(M,p_3^*(A))$ is the strain tensor induced at point M by the loading described by Eq. (1). By varying the position of A on Γ and by applying superposition principle with respect to integration point M, normal displacement in every point of the boundary can be assessed.

The second term in Eq. (2), which is expressed as a volume integral, represents the residual part of displacement, namely the deflection that would persist after unloading elastically the considered body. Knowledge of normal residual displacement allows solving the elastic-plastic contact problem as a purely elastic problem with a modified initial contact geometry. A level of iteration, corresponding to solution of elastic contact, is therefore required for the mutual adjustment between contact pressure and surface normal displacement.

Betti's reciprocal theorem is also applied to assess stress state in the half-space, in the presence of plastic strains. As shown in the following section, knowledge of stress state and of hardening state of the elastic-plastic material allows for computation of plastic strain increment, when a new loading increment is applied leading to further yielding. Again, two independent loads are considered, leading to two independent states (u,ε,σ) and $(u^{**},\varepsilon^{**},\sigma^{**})$, the latter corresponding to a unit load applied along the direction of $\vec{x}_k$, in a point B inside the half-space:

$$p_k^*(M)=\begin{cases}0, & M\neq B;\\ \left(dx_1dx_2dx_3\right)^{-1}, & M=B,\end{cases} \tag{3}$$

The following equation yields from the general form of Betti's reciprocal theorem:

$$u_k(B)=2\mu\int_{\Omega_p}\varepsilon_{ij}^p(M)\varepsilon_{kij}^{**}(M,B)d\Omega+\int_{\Gamma_C}u_{3k}^{**}(M,B)p_3(M)d\Gamma. \tag{4}$$

Here, $u_{3k}^{**}(M,B)$ and $\varepsilon_{kij}^{**}(M,B)$ are the displacement along direction of $\vec{x}_3$ and the ij strain tensor component respectively, induced at point M in the half-space by the unit load applied at point B along the direction of $\vec{x}_k$. By varying the position of B in Ω and by applying superposition principle with respect to integration point M, displacements in every point of the body can be assessed.

Eq. (4) suggests that stresses have an "elastic" part, σ^{pr}, related to contact pressure p_3, which is expressed as a surface integral over Γ_C, and a residual part, σ^r, expressed as a volume integral over plastic region Ω_p. The term "elastic" in the previous statement can be misleading, as all stresses are elastic, but σ^{pr} denotes the part of stresses that would vanish if an elastic unloading would occur. This stresses are related to contact pressure, as opposed to residual stresses σ^r, which are linked to the plastic region Ω_p, and would persist after elastic unloading. If $M_{ijk\ell}$ is the stiffness tensor from Hooke's law, the following equations hold:

$$\sigma_{ij}^{pr}=M_{ijk\ell}\left(\frac{1}{2}(u_{k,\ell}^{pr}+u_{\ell,k}^{pr})\right),\quad u_k^{pr}(B)=\int_{\Gamma_C}u_{k3}^{**}(M,B)p_3(M)d\Gamma, \tag{5}$$

$$\sigma_{ij}^{r}=M_{ijk\ell}\left(\frac{1}{2}(u_{k,\ell}^{r}+u_{\ell,k}^{r})-\varepsilon_{k\ell}^{p}\right),\quad u_k^{r}(B)=2\mu\int_{\Omega_p}\varepsilon_{ij}^p(M)\varepsilon_{kij}^{**}(M,B)d\Omega. \tag{6}$$

A single comma in the subscript denotes the derivative with respect to the corresponding direction: $u_{i,j}=\partial u_i/\partial x_j$.

Resulting equations (2) and (4) suggest elastic-plastic contact problem split in an "elastic" and a residual part. As shown in the following sections, the elastic part comprises the static force equilibrium, interference equation, and complementarity conditions, while the residual part expresses the plastic strain increment and plastic zone contribution to surface normal displacement and to stress state in the elastic-plastic body.

However, the two subproblems cannot be solved independently, as residual displacement, computed in the residual part, enters interference equation in the elastic part, while contact

stress, assessed in the elastic subproblem, is needed to find the plastic strain increment in the residual part.
Analytical resolution of resulting model is available for neither the elastic, nor the residual part, as integration domains, namely boundary region with tractions and plastic strain volume respectively, not known a priori, are arbitrarily shaped. Therefore, numerical approach is preferred.
The principle of numerical approach consists in considering continuous distributions as piece-wise constant on the cells of a three-dimensional grid imposed in a volume enveloping integration domains. Continuous integration in the analytical model of the elastic-plastic contact model is replaced by multi-summation of elementary cells individual contributions. As these multi-summation operations are in fact convolution and/or correlation products, spectral methods are applied to speed up the computation.

3. Numerical solution of the elastic part

The numerical model of the elastic part is obtained from that corresponding to a normal elastic contact problem completed with the residual term, which is superimposed into the interference equation.
Numerical resolution of elastic contact problem relies on considering continuous distributions as piecewise constant on the elements of a rectangular mesh imposed in the common plane of contact and including the contact area. This approach allows transforming the integral contact equation, for which analytical solutions exists only in a few cases, in a linear system of equations, having nodal pressure as unknowns.
Kalker and van Randen, (Kalker & van Randen, 1972), reformulated the elastic contact problem as a problem of minimization, where the unknown contact area and pressure distribution are those who minimize the total complementary energy, under the restrictions that pressure is positive on the contact area and there is no interpenetration. This formulation finally reduces to solving a set of equations and inequalities which have to be satisfied simultaneously:

$$h(i,j) = hi(i,j) + u^{pr}(i,j) - \omega \ , (i,j) \in D \tag{7}$$

$$h(i,j) = 0, p(i,j) > 0 \ , (i,j) \in A \tag{8}$$

$$h(i,j) > 0, p(i,j) = 0 \ , \ (i,j) \in D - A \tag{9}$$

$$\Delta \sum_{(i,j)\in A} p(i,j) = W \tag{10}$$

with: h – the gap between the deformed contact surfaces; hi – the initial gap (without loading); u^{pr} – the composite displacements of the contact surfaces, due to contact pressure; ω - rigid-body approach; W – the load transmitted through contact; A - digitized contact area; D - digitized computational domain. A set of two integers (i,j) is used in the numerical model instead of continuous coordinates x_i to denote patch position in the grid.
This numerical formulation cannot predict singularities in the computed fields, as it employs values averaged over the elementary patches, but allows for the use of influence coefficients based methods. The most efficient approach in solving the system (7)-(10)

employs a modified conjugate gradient method (CGM), originally proposed by Polonsky and Keer, (Polonsky & Keer, 1999). This algorithm has two main advantages over other minimization methods. Firstly, convergence is assured, as there is proof of convergence for the CGM, and the rate of convergence is superlinear. Theory states, (Shewchuk, 1994), that CGM should converge in a number of iterations equal to the number of non-nil unknowns, namely the numbers of cells in contact. In practice, a much faster convergence was observed for smooth contact geometries. Secondly, the algorithm allows for imposing additional restrictions in the course of CG iterations. This means contact area is iterated during pressure correction, based on non-adhesion, Eq. (8), and non-penetration principles, Eq. (9). The force balance condition, Eq. (10), is also imposed to correct the pressure distribution. This eliminates the need for additional nested loops, which were present in most contact solvers prior to this approach.

Convolution product is used to derive the answer of a linear elastic system subjected to an input, when the unit impulse response, also referred to as the Green function, is known. For contact problems, the response of an elastic isotropic half-space to a unit concentrated force applied on the boundary is known from the Boussinesq and/or Cerruti fundamental solutions. The product of this solution (or Green function) with a shape function, as defined in (Liu et al., 2000), yields the influence coefficient (IC), which expresses contribution of an element of the grid into another. Superposition principle is then applied, implying summation of individual contributions over all grid elements. This multi-summation process, which is in fact a convolution product, is very time-consuming, being of order $O(N^2)$ for a grid with N elementary patches.

In order to circumvent this limitation, the solution currently applied is to compute the convolution in the frequency domain, according to convolution theorem, thus reducing the computational effort to $O(N \log N)$. An important issue when using discrete cyclic convolution to assess continuous linear convolution is the periodization of the problem, which induce the so called periodicity error, (Liu et al., 2000). If the Green function is known in the time-space domain, the Discrete Convolution Fast Fourier Transform (DCFFT) technique proposed by these authors, (Liu et al., 2000), eliminates completely the periodicity error, as discrete cyclic convolution approaches the linear continuous convolution the way quadrature estimates continuous integral.

The implemented algorithm for solving numerically the elastic contact problem, described in detail in (Spinu et al., 2007), can be summarized in the following steps:

1. Acquire the input: contact geometry, elastic properties of the contacting materials, normal load transmitted through contact.
2. Establish the computational domain, D. For non-conforming contact problems, Hertz contact area usually makes a good guess value. If during pressure iterations, current contact area is not kept inside computational domain, namely $A^{(k)} \not\subset D$, the algorithm should be restarted with a new D.
3. Establish grid parameters, based on available computational resources.
4. Choose the guess value for pressure, $\mathbf{p}^{(0)}$ and the imposed precision *eps* for the conjugate gradient iteration. According to (Polonsky & Keer, 1999), the latter should be correlated with the number of grids.
5. Start the conjugate gradient loop. Compute surface normal displacement field as a convolution between influence coefficients matrix $\mathbf{K}$ and current pressure $\mathbf{p}^{(k)}$, using DCFFT for computational efficiency: $\mathbf{u}^{(k)} = \mathbf{K} \otimes \mathbf{p}^{(k)}$, where symbol "$\otimes$" is used to denote two-dimensional discrete cyclic convolution.

6. Compute the gap distribution, corresponding to residual in CG formulation, using Eq. (7) with a vanishing rigid body approach ω: $h^{(k)}(i,j) = hi(i,j) + u^{(k)}(i,j), (i,j) \in D$. In order to compensate for the disregarding of ω (which is unknown), $\mathbf{h}^{(k)}$ is normalised by its mean value on the current contact area $A^{(k)}$.
7. Compute the descent direction $d^{(k)}(i,j)$ in the CG algorithm.
8. Compute the length of the step $\alpha^{(k)}$ to be made along minimization direction: $\mathbf{t}^{(k)} = \mathbf{K} \otimes \mathbf{d}^{(k)}$, $\alpha = \mathbf{h}^{(k)}\mathbf{d}^{(k)}\left(\mathbf{t}^{(k)}\mathbf{d}^{(k)}\right)^{-1}$. For consistence with gap correction in step 6, $\mathbf{t}^{(k)}$ is also normalized by its mean value.
9. Adjust nodal pressures: $p^{(k+1)}(i,j) = p^{(k)}(i,j) + \alpha d^{(k)}(i,j)$.
10. Impose complementarity conditions. Cells with negative pressure are excluded from current contact area $A^{(k)}$, and the corresponding nodal pressures are set to zero. Cells with negative gap re-enter $A^{(k)}$, and the corresponding pressures are adjusted according to step 9.
11. Verify convergence criterion: $\left|\mathbf{p}^{(k+1)} - \mathbf{p}^{(k)}\right| \le eps$.

The model was enhanced to allow for eccentric loading of conforming contacts by these authors, (Spinu & Diaconescu, 2008), who imposed an additional Newton-Raphson iterative level to allow for rotation of common plane of contact. Later on, Spinu (Spinu, 2008) further improved the algorithm, by suppressing the outer iterative level and by imposing a correction of tilting angles of contact common plane during CG iterations.

4. Numerical solution of the residual part

4.1 Plastic zone contribution to surface displacement

The residual part is also reformulated numerically, by imposing digitized plastic strain distribution and finite load increments. As the region of plastic strains Ω_p can be arbitrarily shaped, the integrals in Eq. (2) can only be computed numerically. The numerical formulation is based on dividing Ω_p in a set of N cuboids of elementary volume Ω_c, having uniform plastic strains in each elementary cuboid. Consequently, the continuous distribution of $\boldsymbol{\varepsilon}^p$ in Ω_p is assumed as piece-wise constant and Ω_p is substituted by a set of elementary cuboids Ω_{pn}. With this formulation, the residual displacement can be expressed as the sum of contributions of all elementary cuboids in Ω_{pn}:

$$u_3^r(A) = 2\mu \sum_{k=1}^{N} \varepsilon_{ij}^p(k) \int_{\Omega_c} \varepsilon_{3ij}^*(k,A), \tag{11}$$

or, by indexing the cuboids with a set of three integers, and by denoting the cuboid sides with Δ_1, Δ_2 and Δ_3:

$$u_3^r(i,j,0) = 2\mu \sum_{(\ell,m,n)\in\Omega_{pn}} \left\{ \varepsilon_{\zeta\xi}^p(\ell,m,n) \times \int_{x_3(n)-\Delta_3/2}^{x_3(n)+\Delta_3/2} \int_{x_2(m)-\Delta_2/2}^{x_2(m)+\Delta_2/2} \int_{x_1(\ell)-\Delta_1/2}^{x_1(\ell)+\Delta_1/2} \varepsilon_{3\zeta\xi}^*\left(\ell - x_1(i), m - x_2(j), n\right) dx_1\, dx_2 dx_3 \right\}. \tag{12}$$

The tensor ε_3^*, representing strains due to a unit concentrated force applied on surface boundary, is known from Boussinesq fundamental solutions, (Boussinesq, 1969), which represent, in terms of spectral methods, the corresponding Green functions. In order to compute the influence coefficients, functions d_{ii} are defined as primitives of functions $2\mu\varepsilon_{3ii}^* = \mu(u_{3i,i}^* + u_{3i,i}^*)$ with respect to directions of $\vec{x}_1, \vec{x}_2$ and of $\vec{x}_3$, and functions d_{ij}, $i < j$, as primitives of $2\mu\left(\varepsilon_{3ij}^* + \varepsilon_{3ji}^*\right) = 2\mu(u_{3i,j}^* + u_{3j,i}^*)$ with respect to the same directions. The influence coefficients can then be computed according to the formulas given in (Spinu, 2009).

Eq. (12) written with respect to indices of elementary cells takes the following form:

$$u_3^r(i,j,0) = \sum_{(\ell,m,n)\in\Omega_{pm}} \varepsilon_{\zeta\xi}^p(\ell,m,n) D_{\zeta\xi}(\ell - i, m - j, n), \tag{13}$$

with summation over $\zeta, \xi = 1,2,3$, $\zeta \le \xi$. If expression $D_{\zeta\xi}(i - \ell, j - m, n)$ is used in relation (13) instead of $D_{\zeta\xi}(\ell - i, m - j, n)$, namely the point of integration and the point of observation are interchanged, Eq. (13) takes the following form:

$$u_3^r(i,j,0) = \sum_{(\ell,m,n)\in\Omega_{pm}} D_{\zeta\xi}(i - \ell, j - m, n) \varepsilon_{\zeta\xi}^p(\ell,m,n), \tag{14}$$

which represents a discrete cyclic convolution with respect to directions of $\vec{x}_1$ and of $\vec{x}_2$. Efficient computation for this product is available through DCFFT, (Liu et al., 2000).

4.2 Plastic zone contribution to stress state

The problem of residual stresses due to plastic zone in elastic-plastic contact can be treated in the more general frame of the so called "inclusion problem". Eigenstrains such as plastic strains, misfits strains, thermal expansion or phase transformation, generate a linear elastic stress field in an isotropic half-space. Usually, assessment of this field, also referred to as the inclusion problem, is performed using a problem decomposition method originally suggested by Chiu, (Chiu, 1978). Although inclusion problem has received a great deal of attention in the last four decades, (Mura, 1988), closed form solutions exist only in a few cases of simple, regular shapes, such as spherical or cuboidal eigenstrains. In elastic-plastic contact modeling, these limiting assumptions are not met, thus imposing the use of numerical approach.

The problem of residual stresses arising in elastic-plastic contact was solved by Mayeur, (Mayeur, 1995), for the two-dimensional rough contact. The three-dimensional case was solved by Jacq, (Jacq, 2001), using Chiu's problem decomposition, (Chiu, 1978). These authors, (Jacq et al., 2002), used two-dimensional fast Fourier transform algorithms to efficiently compute the arising convolution products. Wang and Keer, (Wang & Keer, 2005), used a similar approach in studying residual stresses arising in elastic-plastic contact with hardening behavior. They stated that two-dimensional DCFFT should be applied in residual stress computation.

An alternative to Chiu's problem decomposition was advanced by Liu and Wang, (Liu & Wang, 2005), based on Mindlin and Cheng's results, (Mindlin & Cheng, 1950), involving derivatives of four key integrals. They also advanced an efficient algorithm to compute

correlation products using convolution theorem, called Discrete Correlation Fast Fourier Transform (DCRFFT).

Jin, Keer, and Wang, (Jin et al., 2008), suggested that, in order to achieve a better computational efficiency, convolution and correlation should be used together, in a hybrid algorithm. They presented some comparative results obtained using both two-dimensional and three-dimensional spectral algorithms, proving that the latter reduces dramatically the CPU time and memory requirements, allowing for finer grids.

The problem of elastic fields due to arbitrarily shaped inclusions in an elastic half-space was also treated by these authors, (Zhou et al., 2009). Although Chiu's problem decomposition is employed, influence coefficients for imposing the pressure-free surface condition are not derived explicitly, as stresses due to spurious pressure on the boundary are not expressed as functions of existing eigenstrains.

Mura, (Mura, 1968), stated that, in the presence of initial strains, a finite body with a traction-free surface can be treated as an infinitely extended body, if equal and opposite normal and shear stresses are applied on the boundary, compensating for the ones corresponding to the full space solution. Consequently, the method suggested by Chiu, (Chiu, 1978) consists in applying superposition principle to elastic states (b), (c), and (d) in Fig. 1, whose summation yields the elastic state of the original problem (a).

Eigenstrains in state (b) are identical to those of the original problem (a), while in state (c), the cuboid is the mirror image of the original one with respect to half-space boundary. Eigenstrains in state (c) are chosen such as shear tractions induced by states (b) and (c) cancel each-other on the half-space boundary:

$$\boldsymbol{\varepsilon}^{pm} = \boldsymbol{\varepsilon}^{p} \text{ , except for } \varepsilon_{13}^{pm} = -\varepsilon_{13}^{p} \text{ , and } \varepsilon_{23}^{pm} = -\varepsilon_{23}^{p} \text{ ,} \tag{15}$$

leading to a spurious normal traction (or pressure) depicted by state (d). Consequently, in order to simulate the traction-free boundary condition, solution of state (d) should be extracted from summation of solutions corresponding to states (b) and (c).

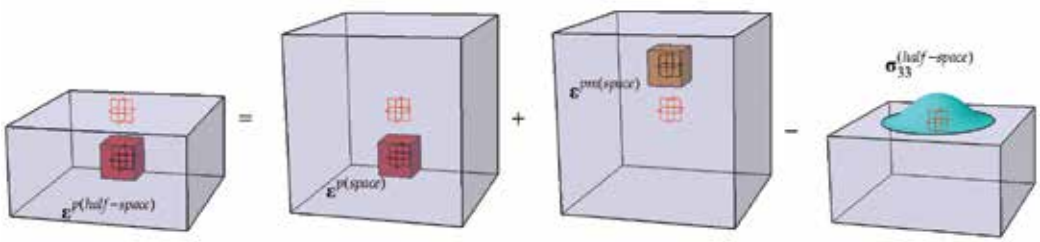


Fig. 1. Inclusion problem decomposition: a. cuboidal inclusion in elastic half-space; b. cuboidal inclusion in infinite elastic space; c. an image counterpart in infinite space; d. a half-space with a pressure distribution

A uniformly-spaced rectangular grid is established in a cuboidal domain including the arbitrarily shaped plastic zone. According to superposition principle, problem solution is obtained by superimposing the solution of each cuboidal inclusion. If the grid is uniformly spaced, the number of different influence coefficients to be computed is reduced to the number of different distances between cell control points. This allows reformulation of multi-summation operation as a discrete convolution, which can be evaluated efficiently in the frequency domain, according to convolution and/or correlation theorems.

Plastic strains are assumed constant in every elementary cell, but otherwise can vary along computational domain. The solution for a cuboidal inclusion of constant eigenstrains in an infinite space, namely the IC, is needed.

The first closed form solution for the ICs required to assess states (b) and (c) in Fig. 1 was advanced by Chiu, (Chiu, 1977). A Cartesian coordinate system (x_1', x_2', x_3') is attached to the centre of the cuboid. In the presence of plastic strains ε_{ij}^p, displacements u_i are related to strains by the strain-displacement equations:

$$\varepsilon_{ij}^e + \varepsilon_{ij}^p = \frac{1}{2}\left(u_{i,j} + u_{j,i}\right), \tag{16}$$

where $\boldsymbol{\varepsilon}^e$ is the elastic component of strains. By substituting ε_{ij}^e into the constitutive equation (Hooke's law), one can find the stresses induced by the eigenstrains ε_{ij}^p. The gradients of displacements needed in Eq. (16) were obtained by Chiu, (Chiu, 1977), using the Galerkin vector:

$$2\mu u_{i,q}(x_1', x_2', x_3') = \frac{1}{8\pi^3}\sum_{m=1}^{8}(-1)^m \left[\begin{array}{l} \dfrac{1-2\nu}{1-\nu}\lambda\varepsilon_{kk}^p D_{,iqnn}(\mathbf{c}_m) + 4\mu\varepsilon_{ij}^p D_{,jqnn}(\mathbf{c}_m) - \\ -\dfrac{2\mu}{1-\nu}\varepsilon_{nj}^p D_{,iqnj}(\mathbf{c}_m) \end{array}\right], \tag{17}$$

where μ and λ are Lamé's constants, $\mathbf{c}_m$, $m = \overline{1,8}$ are the eight vectors linking the corners of the cuboid to the observation point, and $D(\mathbf{c}_m)$ is a function whose fourth derivates with respect to coordinates x_j' are obtained by circular permutation in one of four categories, $D_{,1111}$, $D_{,1112}$, $D_{,1122}$ and $D_{,1123}$, given in (Chiu, 1977). Einstein summation convention is employed in Eq. (17).

Summation of elastic fields induced by $\boldsymbol{\varepsilon}^p$ and $\boldsymbol{\varepsilon}^{pm}$ in a coordinate system with the origin on the half-space boundary yields the following equation:

$$\begin{aligned} \sigma_{ij}^{(space)}(x_1, x_2, x_3) = A_{ijk\ell}(x_1 - x_1', x_2 - x_2', x_3 - x_3')\varepsilon_{k\ell}^p(x_1', x_2', x_3') + \\ A_{ijk\ell}(x_1 - x_1', x_2 - x_2', x_3 + x_3')\varepsilon_{k\ell}^{pm}(x_1', x_2', -x_3'). \end{aligned} \tag{18}$$

where (x_1, x_2, x_3) is the observation point and (x_1', x_2', x_3') the source point (the control point of the elementary cuboid having uniform plastic strains).

As all distributions are assumed piece-wise constant, it is convenient to index the collection of cuboids by a sequence of three integers ranging from 1 to N_1, N_2 and N_3 respectively, with $N = N_1 N_2 N_3$, and to express all distributions as functions of these integers instead of coordinates.

After superimposing the individual contributions of all cuboids, Eq. (18) becomes:

$$\begin{aligned} \sigma_{\xi\zeta}^{r(space)}(i,j,k) = \sum_{\ell=1}^{N_1}\sum_{m=1}^{N_2}\sum_{n=1}^{N_3} A_{\xi\zeta\varsigma\gamma}(i-\ell, j-m, k-n)\varepsilon_{\varsigma\gamma}^p(\ell,m,n) + \\ \sum_{\ell=1}^{N_1}\sum_{m=1}^{N_2}\sum_{n=1}^{N_3} A_{\xi\zeta\varsigma\gamma}(i-\ell, j-m, k+n)\varepsilon_{\varsigma\gamma}^p(\ell,m,n), \end{aligned} \tag{19}$$

which expresses the stress field induced in infinite space at cell (i,j,k) by all cuboids of uniform eigenstrains (ℓ,m,n) and by their mirror images.
Based on this development, the spurious normal traction induced on the half-space boundary, $\sigma_{33}^{(half-space)}$, needed to solve the state (d) in Fig. 1, can be expressed:

$$\sigma_{33}^{(half-space)}(i,j)=\sigma_{33}^{r(space)}(i,j,0)=\sum_{\ell=1}^{N_1}\sum_{m=1}^{N_2}\sum_{n=1}^{N_3}A_{33\varsigma\gamma}(i-\ell,j-m,-n)\varepsilon_{\varsigma\gamma}^{p}(\ell,m,n)+\sum_{\ell=1}^{N_1}\sum_{m=1}^{N_2}\sum_{n=1}^{N_3}A_{33\varsigma\gamma}(i-\ell,j-m,n)\varepsilon_{\varsigma\gamma}^{p}(\ell,m,n), \tag{20}$$

The stress induced in the half-space by this fictitious traction can then be computed:

$$\sigma_{\zeta\xi}(i,j,m)=\sum_{k=1}^{N_1}\sum_{\ell=1}^{N_2}Q_{\zeta\xi}(i-k,j-\ell,m)\sigma_{33}^{(half-space)}(k,\ell). \tag{21}$$

The influence coefficients Q_{ij}, (Liu and Wang, 2002), result from integration of Boussinesq formulas over elementary grid cell with respect to directions of $\vec{x}_1$ and $\vec{x}_2$. The product in Eq. (21) is a two-dimensional convolution with respect to directions of $\vec{x}_1$ and $\vec{x}_2$, which can be computed efficiently with DCFFT algorithm.
Finally, the solution for the stress due to arbitrarily shaped eigenstrains in an elastic isotropic half-space results from superposition of solutions (19) and (21).
The two terms in Eq. (19) imply multi-summation over three dimensions, as both source and observation domains are three-dimensional. Computation of these distributions by direct multiplication method (DMM) or even by two-dimensional DCFFT is very time-consuming, therefore a non-conventional approach is required. The first term in Eq. (19) is a three-dimensional convolution, while the second term is a two-dimensional convolution with respect to directions of $\vec{x}_1$ and $\vec{x}_2$ and a one-dimensional correlation with respect to direction of $\vec{x}_3$. Liu and Wang, (Liu & Wang, 2005), suggested that correlation theorem, together with convolution theorem, could be used together in a hybrid convolution-correlation multidimensional algorithm.
In the last decade, spectral methods are intensively used in contact mechanics to rapidly evaluate convolution-type products. These authors, (Jacq et al., 2002), applied a two-dimensional fast Fourier transform algorithm to speed up the computation of convolution products arising in Eq. (19). Their approach reduces the computational requirements from $O(N_1^2N_2^2N_3^2)$ in DMM to $O(N_3^2N_1N_2\log N_1N_2)$.
However, using a two-dimensional algorithm to solve a problem which is essentially three-dimensional is an imperfect solution. Therefore, in this work, a three-dimensional spectral algorithm is implemented, capable of evaluating both convolution and hybrid convolution-correlation type products in $O(N_1N_2N_3\log N_1N_2N_3)$ operations. The algorithm, originally advanced in (Spinu & Diaconescu, 2009), is based on the notorious DCFFT technique (Liu et al., 2000).
If the ICs are known in the time/space domain, this algorithm can evaluate the linear convolution by means of a cyclic convolution with no periodicity error. The concepts of "zero-padding" and "wrap-around order", presented in (Liu et al., 2000), can be extended

naturally to the three-dimensional case, and applied to compute the first term in the right side of Eq. (19). However, for the second term, due to positioning of the mirror-image element relative to global coordinate system (linked to half-space boundary), convolution turns to correlation with respect to direction of $\vec{x}_3$. In order to use three-dimensional FFT and convolution theorem to evaluate the convolution-correlation product, the following algorithm is proposed:

1. The influence coefficients $\mathbf{A}$ are computed as a three dimensional array of $N_1 \times N_2 \times 2N_3$ elements, using the formulas derived from Eqs. (16) and (17).
2. The term $\mathbf{A}$ is extended into a $2N_1 \times 2N_2 \times 2N_3$ array by applying zero-padding and wrap-around order with respect to directions of $\vec{x}_1$ and $\vec{x}_2$, as requested by the classic DCFFT algorithm.
3. Plastic strains $\boldsymbol{\varepsilon}^p$ are inputted as a three-dimensional array of $N_1 \times N_2 \times N_3$ elements.
4. The term $\boldsymbol{\varepsilon}^p$ is extended to a $2N_1 \times 2N_2 \times 2N_3$ array by zero-padding in all directions.
5. Elements of $\boldsymbol{\varepsilon}^p$ are rearranged in reversed order with respect to direction of $\vec{x}_3$.
6. The Fourier transforms of $\mathbf{A}$ and $\boldsymbol{\varepsilon}^p$ are computed by means of a three-dimensional FFT algorithm, thus obtaining the complex arrays $\hat{\mathbf{A}}$ and $\hat{\boldsymbol{\varepsilon}}^p$, where $(\hat{g})$ is used to denote the discrete Fourier transform of any time/space array g.
7. The spectral array of residual stresses is computed as element-by-element product between convolution terms: $\hat{\sigma}^{r(space)} = \hat{A} \cdot \hat{\varepsilon}^p$.
8. The time/space array of residual stresses is finally obtained by means of an inverse discrete Fourier transform: $\sigma^{r(space)} = IFFT(\hat{\sigma}^{r(space)})$.
9. The terms in the extended domain are discarded, thus keeping the terms $N_1 \times N_2 \times N_3$ of $\sigma^{r(space)}$ as output.

Domain extension with respect to directions of $\vec{x}_1$ and $\vec{x}_2$ in step 2 is required by the DCFFT technique, and no additional treatment is needed to evaluate the corresponding discrete cyclic convolutions. On the other hand, according to discrete correlation theorem, (Press et al., 1992), a correlation product can be evaluated as a convolution between one member of the correlation and the complex conjugate of the other. Therefore, DCFFT can be applied with respect to direction of $\vec{x}_3$ too, if the second term, namely the plastic strains array, is substituted by its complex conjugates in the frequency domain. The fastest way to achieve this is to rearrange the terms of $\boldsymbol{\varepsilon}^p$, as indicated in step 4. Indeed, when FFT is applied on a series of real terms g, thus obtaining $\hat{g}$, one can obtain its complex conjugate $\hat{g}^*$, simply by reading g in reversed order. This remarkable property allows for combining convolutions and correlations products with respect to different directions in a hybrid algorithm. By applying three-dimensional FFT, the computational effort for solving the inclusion problem in infinite, elastic and isotropic space is reduced considerably, from $O(N_3^2 N_1 N_2 \log N_1 N_2)$ in Jacq's approach to $O(N_1 N_2 N_3 \log N_1 N_2 N_3)$ operations for the newly proposed algorithm.

The following step is to compute the stress state induced in the half-space by spurious normal traction $\sigma_{33}^{(half-space)}$. In existing formulations, (Chiu, 1978; Jacq, 2001), this stresses are expressed explicitly as functions of plastic strains ε_{ij}^p. This rigorous formulation results in increased model complexity. It also has the disadvantage of limiting the application of spectral methods to two-dimensional case. However, if the analysis domain is large enough, one can assume that the normal traction induced on the half-space boundary vanishes outside the computational domain. Therefore, the corresponding elastic state (d) is due to term $\sigma_{33}^{(half-space)}$ alone. With this assumption, computation of elastic state (d) is

reduced to the problem of a stress state induced in an elastic isotropic half-space by an arbitrarily, yet known, pressure (or normal traction). Solution of this problem is readily available, as corresponding Green functions are known from Boussinesq fundamental solutions.

The resulting computational advantage is more effective when using the newly proposed algorithm as part of an elastic-plastic contact code. Indeed, influence coefficients Q_{ij} needed to assess stresses induced by pressure are shared with the elastic contact code. They are computed and stored as a $N_1 \times N_2 \times N_3$ array. In Jacq's formulation, N_3 arrays, each having $N_1 \times N_2 \times N_3$ terms, are needed, because influence coefficients needed to impose free surface relief depend explicitly on both source and computation point depths. This double dependence also limit the use of spectral methods to two dimensions, thus being of order $O(N_3^2 N_1 N_2 \log N_1 N_2)$, corresponding to N_3^2 two-dimensional DCFFTs in layers of constant depth.

In the simplified formulation advanced in this paper, as source domain (namely pressure domain) is only two-dimensional, as opposed to plastic zone, which is three-dimensional, the computational order is decreased to $O(N_1 N_2 N_3 \log N_1 N_2)$ operations, corresponding to N_3 two-dimensional DCFFTs in layers of constant depth.

The method for imposing the pressure-free condition assumes that spurious normal tractions on the half-space boundary vanish outside computational domain. This assumption requires a larger computational domain in order to minimize truncation errors. When simulating concentrated elastic-plastic contacts, plastic region is usually located under the central region of the contact area, occupying a hemispherical domain. Therefore, the newly proposed method is well adapted to this kind of problems.

As inclusion problem has to be solved repeatedly in an elastic-plastic contact simulation, the overall computational advantage is remarkable, allowing for finer grids or smaller loading steps to reduce discretization error.

4.3 Plastic strain increment assessment

According to general theory of plasticity, plastic flow occurrence can be described mathematically with the aid of a yield function, assessing the yield locus in the multidimensional space of stress tensor components. If von Mises criterion is used to assess stress intensity, this function can be expressed as:

$$f(e^p) = \sigma_{VM} - \sigma_Y(e^p), \tag{22}$$

where e^p denotes the effective accumulated plastic strain, $e^p = \sqrt{2\varepsilon_{ij}^p \varepsilon_{ij}^p / 3}$, and $\sigma_Y(e^p)$ is the yield strength function. The latter satisfy the relation for the initial yield strength σ_{Y0}:

$$\sigma_Y(0) = \sigma_{Y0}. \tag{23}$$

For elastic-perfectly plastic materials, relation (23) is verified for any value of e^p. However, for metallic materials, more complex models of elastic-plastic behavior are employed, as the isotropic, or the kinematic hardening laws. The isotropic hardening law of Swift,

$$\sigma_Y(e^p) = B(C + e^p)^n, \tag{24}$$

with B,C and n material constants, is used in the current formulation, as it is verified for many metallic materials, (El Ghazal, 1999) and, from a numerical point of view, it has the advantage of being continuously derivable.
The following conditions must be met all the time:

$$f \le 0;\ de^p \ge 0;\ f \cdot de^p = 0\,, \tag{25}$$

with $f = 0$ and $de^p > 0$ corresponding to plastic flow.
According to flow rule, plastic strain increment can be expressed as:

$$d\varepsilon_{ij}^p = de^p \frac{\delta f}{\delta \sigma_{ij}} = de^p \frac{3S_{ij}}{2\sigma_{VM}}\,, \tag{26}$$

where S_{ij} denotes the deviatoric stress tensor.
The algorithm used to derive the plastic strain increment was advanced by Fotiu and Nemat-Nasser, who developed a universal algorithm for integration of elastoplasticity constitutive equations. As stated in (Fotiu & Nemat-Nasser, 1996), the algorithm is unconditionally stable and accurate even for large load increments, as it takes into account the entire non-linear structure of elastoplasticity constitutive equations. These are solved iteratively, via Newton-Raphson numerical method, at the end of each loading step. The yield function f is linearized at the beginning of the load increment, by employing an elastic predictor. This places the predictor (trial) state far outside the yield surface $f = 0$, since elastic-plastic modulus is small compared to the elastic one. The return path to the yield surface is generated by the plastic corrector, via Newton-Raphson iteration. This approach, also referred to as elastic predictor - plastic corrector, is efficient when most of the total strain is elastic. In the fully plastic regime, which occurs usually after the elastic-plastic one, the plastic strain is predominant, thus the return path may require numerous iterations. Thus, linearization at the beginning of the loading step is performed by a plastic predictor, and return path is generated with an elastic corrector.
A yield occurs when von Misses stress exceeds current yield stress, namely when $f > 0$. The elastic domain expands and/or translates to include the new state, namely to verify condition $f = 0$. The actual increment of effective accumulated plastic strain should satisfy, in the plastic zone, equation of the new yield surface:

$$f(e^p + \delta e^p) = 0\,. \tag{27}$$

Here, δe^p denotes the finite increment of effective plastic strain, as defined in (Jacq, 2001). Relation (27) can be considered as an equation in δe^p, which is solved numerically by Newton-Raphson iteration. To this end, yield surface relation is linearized along plastic corrector direction:

$$f(e^p + \delta e^p) = f(e^p) + \delta e^p \frac{\partial f(e^p)}{\partial e^p} = 0\,, \tag{28}$$

yielding the plastic corrector:

$$\delta e^p = -\frac{f(e^p)}{\dfrac{\partial f(e^p)}{\partial e^p}} = \frac{f(e^p)}{\dfrac{\partial \sigma_Y(e^p)}{\partial e^p} - \dfrac{\partial \sigma_{VM}}{\partial e^p}}. \tag{29}$$

For isotropic hardening, the derivate of equivalent von Mises stress with respect to effective accumulated plastic strain was derived by Nélias, Boucly and Brunet, (Nélias et al., 2006), from the general equations presented in (Fotiu & Nemat-Nasser, 1996) for rate-dependent elastoplasticity:

$$\frac{\partial \sigma_{VM}}{\partial e^p} = -3G, \tag{30}$$

where G is the shear modulus, or the μ Lamé's constant.

With these results, the following return-mapping algorithm with elastic predictor - plastic corrector can be formulated:

1. Acquire the state at the beginning of the loading step and impose the elastic predictor. For elastic-plastic contact problems, this is equivalent to solving an elastic loop without imposing any residual displacement increment. Corresponding parameters are identified by an "a" superscript, as opposed to a "b" superscript, used to denote the state at the end of the loading increment: $e^{p(a)}$, $\sigma_Y^{(a)} = \sigma_Y(e^{p(a)})$, $\sigma_{ij}^{(a)} = \sigma_{ij}^{pr(a)} + \sigma_{ij}^{r(a)}$, $\sigma_{VM}^{(a)}$, $f^{(a)} = \sigma_{VM}^{(a)} - \sigma_Y^{(a)}$. These variables also represent the input for the Newton-Raphson iteration. Thus, by using superscripts to denote the Newton-Raphson iteration number, $e^{p(1)} = e^{p(a)}$, $\sigma_Y^{(1)} = \sigma_Y^{(a)}$, $\sigma_{ij}^{(1)} = \sigma_{ij}^{(a)}$, $\sigma_{VM}^{(1)} = \sigma_{VM}^{(a)}$, $f^{(1)} = f^{(a)}$.
2. Start the Newton-Raphson iteration. Compute the plastic corrector according to relations (29) and (30):

$$\delta e^{p(i)} = f^{(i)} \Bigg/ \left(\frac{\partial k(e^{p(i)})}{\partial e^{p(i)}} + 3G \right). \tag{31}$$

3. Use the plastic corrector to adjust model parameters:

$$\sigma_{VM}^{(i+1)} = \sigma_{VM}^{(i)} - 3G\delta e^{p(i)}; \quad e^{p(i+1)} = e^{p(i)} + \delta e^{p(i)}; \quad \sigma_Y^{(i+1)} = \sigma_Y(e^{p(i+1)}); \quad S_{ij}^{(i+1)} = \frac{\sigma_{VM}^{(i+1)}}{\sigma_{VM}^{(1)}} S_{ij}^{(1)}. \tag{32}$$

4. Verify if Eq. (27) is verified to the imposed tolerance eps. If condition

$$\left| f^{(i+1)} \right| = \left| \sigma_{VM}^{(i+1)} - \sigma_Y^{(i+1)} \right| > eps \tag{33}$$

is satisfied, go to step 2. If else, convergence is reached, and the state at the end of the loading step is described by the newly computed parameters: $e^{p(b)} = e^{p(i+1)}$, $\sigma_{VM}^{(b)} = \sigma_{VM}^{(i+1)}$, $S_{ij}^{(b)} = S_{ij}^{(i+1)}$.
5. Compute the plastic strain increment, according to Eq. (26):

$$\delta \varepsilon_{ij}^p = \left(e^{p(b)} - e^{p(a)} \right) \frac{3S_{ij}^{(b)}}{2\sigma_{VM}^{(b)}}. \tag{34}$$

This increment is used to update the plastic zone. The residual parts of displacement and of stress can then be computed, and superimposed to their elastic counterparts.

5. Numerical solution of the elastic-plastic contact problem

Elastic-plastic normal contact problem is solved iteratively based on the relation between pressure distribution and plastic strain, until the latter converges. Plastic strain modifies contact pressure by superposing induced residual surface displacement into the interference equation. Contact pressure, in its turn, contributes to the subsurface stress state, responsible for plastic strain evolution.

Finally, the algorithm proposed for simulation of elastic-plastic contact with isotropic hardening is based on three levels of iteration:

1. The innermost level, corresponding to the residual part, assesses plastic strain increment, based on an algorithm described in the previous section, and the contribution of plastic zone to stress state and surface displacement.
2. The intermediate level adjusts contact pressure and residual displacement in an iterative approach specific to elastic contact problems with arbitrarily shaped contact geometry.
3. The outermost level is related to the fact that, unlike elastic solids, in which the state of strain depends on the achieved state of stress only, deformation in a plastic body depends on the complete history of loading. Plasticity is history dependent, namely current state depends upon all pre-existing states. In this level, the load is applied in finite increments, starting from an intensity corresponding to elastic domain, until the imposed value is reached.

The algorithm for solving one loading step in the elastic-plastic normal contact problem is summarized in Fig. 2.

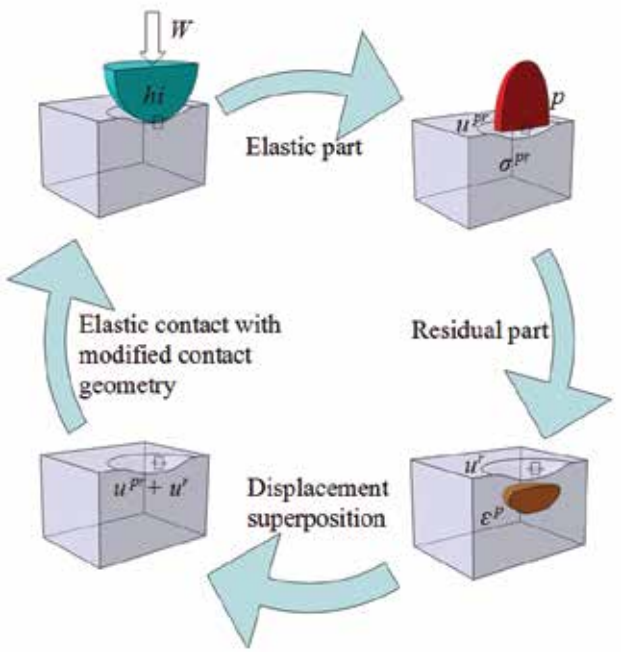


Fig. 2. Elastic - plastic algorithm

Firstly, the elastic problem with modified contact geometry hi is solved, yielding contact area and pressure distribution p. The latter is used to assess elastic displacement field u^{pr} and stress field σ^{pr}. These terms represent the "elastic" part of displacement and of stress, namely that part that is recovered once loading is removed (after contact opening). The stresses induced by pressure are used, together with hardening state parameters, in the residual subproblem, to assess plastic strain increment and to update the achieved plastic zone ε^p. Residual parts of displacement, u^r, and of stresses, σ^r, can then be computed. As opposed to their elastic counterparts, the terms u^r and σ^r express a potential state, that would remain after contact unloading, if no plastic flow would occur during load relief. The total displacement can then be computed, $u^{pr} + u^r$, thus imposing a new interference equation in the elastic subproblem. These sequences are looped until convergence is reached.

The new algorithm for computation of plastic strain increment improves dramatically the speed of convergence for the residual subproblem. The formulation advanced by Jacq, (Jacq, 2001), based on the Prandtl-Reuss algorithm, implies iteration of a tensorial parameter, namely the plastic strain increment, as opposed to the new algorithm, which iterates a scalar, namely the increment of effective accumulated plastic strain. Convergence of the Newton-Raphson scheme is reached after few iterations. As stated in (Fotiu & Nemat-Nasser, 1996), the method is accurate even for large loading increments.

Moreover, Jacq's algorithm is based on the reciprocal adjustment between plastic strain and residual stress increments. Consequently, at every iteration of the residual loop (the innermost level of iteration), it is necessary to express the residual stress increment. Its assessment implies superposition, with both source (integration) and observation domains three-dimensional. Although three-dimensional spectral methods were implemented to speed up the computation, the CPU time and memory requirements remain prohibitively high.

In the new algorithm, residual stresses due to plastic zone needs to be evaluated at every iteration of the elastic loop (the intermediate level of iteration), after plastic zone update with the new plastic strain increment. In other words, residual stress assessment is moved to an upper iterative level, resulting in increased computational efficiency. Consequently, with the same computational effort, a finer grid can be imposed in the numerical simulations, thus reducing the discretization error.

6. Numerical simulations and program validation

In this section, numerical predictions of the newly proposed algorithm are compared with already published results, validating the computer code. The materials of the contacting bodies are assumed to be either rigid (R), or elastic (E), or elastic-plastic (EP), having a behavior described by a power hardening law (Swift), or elastic-perfectly-plastic (EPP). Four types of contacts are considered: R-EP, E-EP, EP-EP with symmetry about the common plane of contact and R-EPP.

Development of plastic region and of residual stresses with application of new loading increments is assessed, and contribution of residual state, which superimpose elastic state induced by contact pressure, is suggested.

Algorithm refinements allow for a fine grid, of $120 \times 120 \times 80$ elementary cells, to be imposed in the computational domain.

6.1 R-EP contact

The contact between a rigid sphere of radius $R = 105 \cdot 10^{-6} m$ and an elastic-plastic half-space is simulated, allowing for comparison with results published by Boucly, Nélias, and Green, (Boucly et al., 2007). Elastic half-space parameters are: Young modulus, $E_2 = 210 GPa$, Poisson's ratio, $\nu_2 = 0.3$. The hardening law of the elastic-plastic material is chosen as a power law (Swift), according to (El Ghazal, 1999), Eq. (24), with e^p the effective accumulated plastic strain, expressed in microdeformations, and the following parameters: $B = 1,280 MPa$, $C = 30$, $n = 0.085$.

The contact is loaded incrementally up to a maximum value of $W = 0.65N$, for which the purely elastic model (Hertz) predicts a contact radius $a_H = 6.053 \mu m$ and a hertzian pressure $p_H = 8,470\ MPa$.

Dimensionless coordinates are defined as ratios to a_H, $\bar{x}_i = x_i / a_H$, and dimensionless pressure or stresses as ratios to p_H. The computational domain is a rectangular cuboid of sides $L_1 = L_2 = 3a_H$, $L_3 = 1.6a_H$, which is dicretized with the following parameters: $N_1 = N_2 = 120$, $N_3 = 80$ elementary grid cells. Due to the fact that problem is axisymmetric, three dimensional distributions are depicted in the plane $x_2 = 0$ only.

Pressure profiles predicted by the numerical program for six loading levels corresponding to elastic-plastic domain are depicted in Fig. 3. Hertz pressure corresponding to maximum load is also plotted for reference.

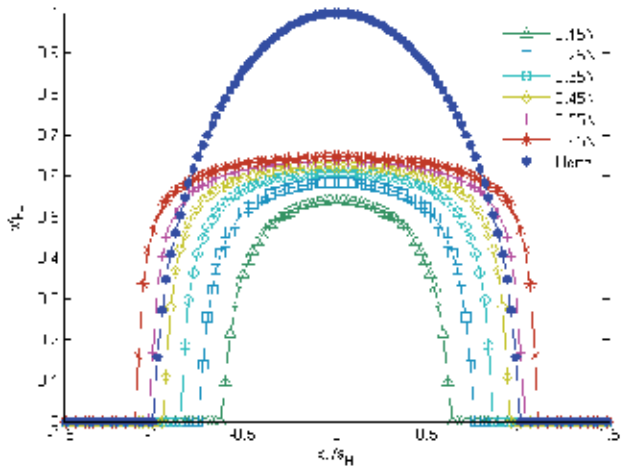


Fig. 3. Pressure profiles in the plane $x_2 = 0$, various loading levels

Elastic-plastic pressure distributions appear flattened compared to the purely elastic case. At the end of the loading loop, a central plateau of uniform pressure can be observed in the vicinity of $6.5 p_H$. This limitation of contact pressure results in an increased elastic-plastic contact radius, compared to its elastic counterpart, a_H.

The same distributions were obtained by Jacq et al., (Jacq et al., 2002), by Boucly, Nélias, and Green, (Boucly et al., 2007), using load driven (ld) or displacement driven (dd) formulations, and also by Benchea and Cretu, (Benchea & Cretu, 2008), using finite element analysis (FEA).

Initiation of plastic flow occurs on the contact axis, where von Mises equivalent stress firstly exceeds initial yield strength. With application of new loading increments, plastic zone

expands to a hemispherical domain, Fig. 4, while material hardening state is modified according to Eq. (24).

Toward the end of the loading cycle, the plastic core approach peripherally the free surface, enveloping an elastic core. Evolution of maximum effective accumulated plastic strain with loading level is presented in Fig. 5.

The model assumes elastic and plastic strains are of the same order of magnitude, corresponding to elastic-plastic range. As plastic strains are small, usually less than 2%, they can be considered small strains and can be superimposed to their elastic counterparts. This approach cannot be applied to larger plastic strains, corresponding to fully plastic range, solution of this scenario requiring FEA.

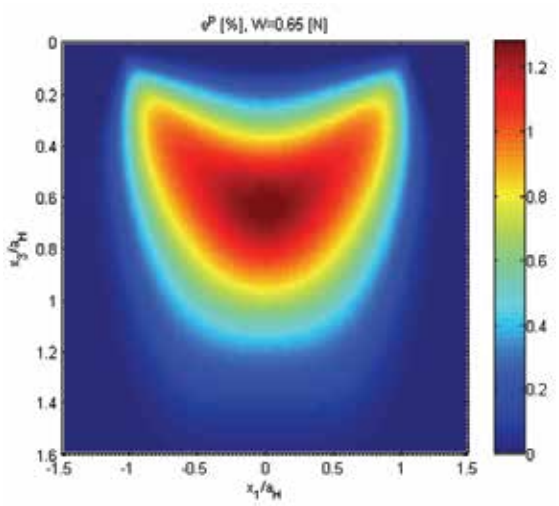


Fig. 4. Effective accumulated plastic strain at $W = 0.65N$

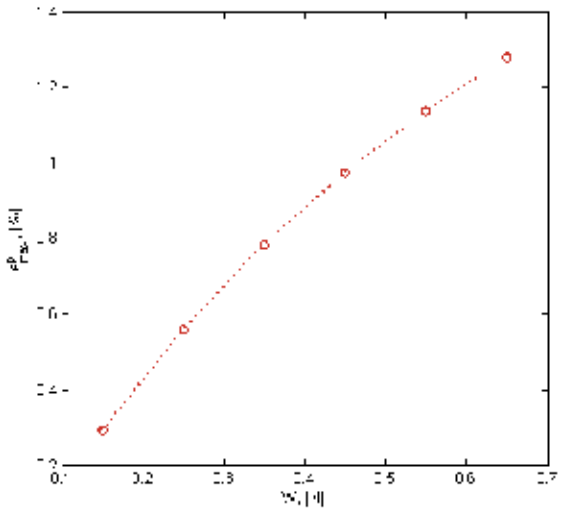


Fig. 5. Maximum effective accumulated plastic strain versus loading level

Plastic strains induce residual stresses, namely elastic stresses that would persist after elastic unloading. These stresses superimpose the ones induced by contact pressure. The resulting state generates further plastic strain if stress intensity exceeds yield strength. Consequently, an accurate estimation of stress field in the elastic-plastic body is essential to plastic strain increment prediction.

Figures 6 and 7 depict distributions of equivalent von Mises contact stress (stress induced by contact pressure) and total stress in the elastic-plastic half-space. Residual stress intensity, Fig. 8, is one order of magnitude smaller than equivalent contact stress. Comparison of distributions depicted in Figs. 6 and 7, using the same scale, suggests that residual stress reduces peaks in contact stress intensity, thus making the resulting field more uniform. This behavior is also suggested by the curves traced in Fig. 9. Maximum intensity of contact stress increase more rapidly than the maximum of the total field, due to contribution of residual stress. Consequently, residual stresses, which represent material response to plastic flow, act to impede further plastic yielding.

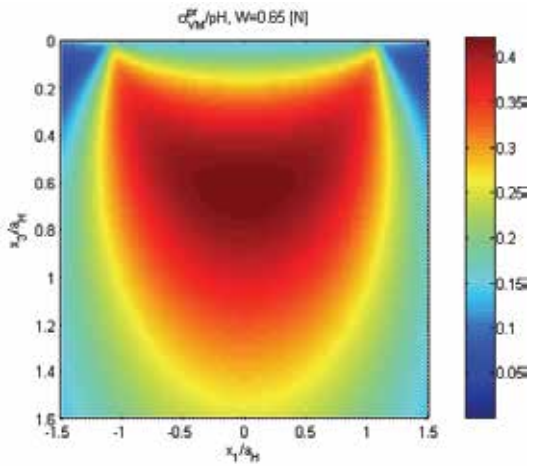


Fig. 6. Von Mises stress induced by contact pressure

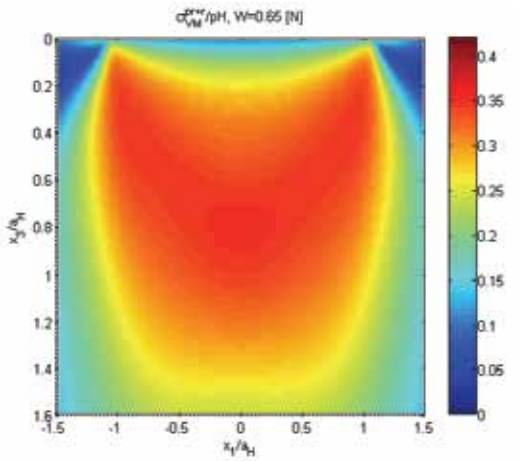


Fig. 7. Maximum intensities of stress fields versus loading level

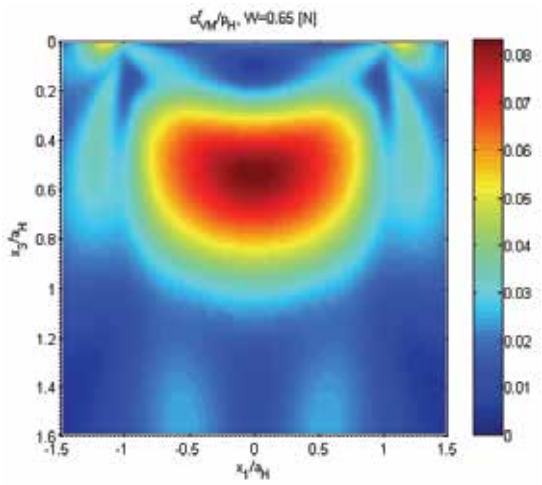


Fig. 8. Von Mises residual stress

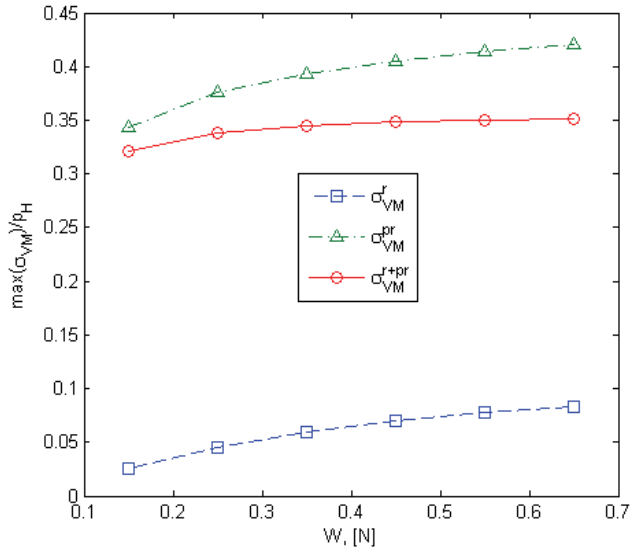


Fig. 9. Total (contact and residual) Von Mises stress in the elastic-plastic body

Profiles of residual prints corresponding to the same six loading levels are depicted in Fig. 10. These profiles show that residual displacement increase contact conformity in investigated non-conforming contact, leading to a more uniform distribution of contact pressure.

The variation of residual print maximum depth with the loading level is presented in Fig. 11. This curve was also obtained experimentally by El Ghazal, (El Ghazal, 1999), numerically by Jacq et al., (Jacq et al., 2002), and using FEA by Benchea and Cretu, (Benchea & Cretu, 2008).

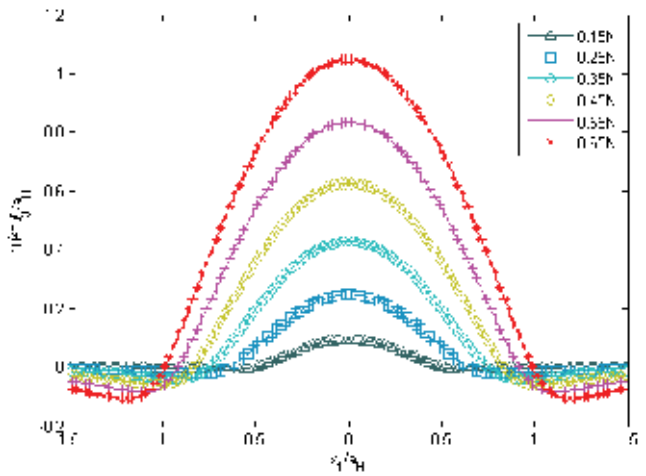


Fig. 10. Residual print profiles in elastic-plastic spherical contact

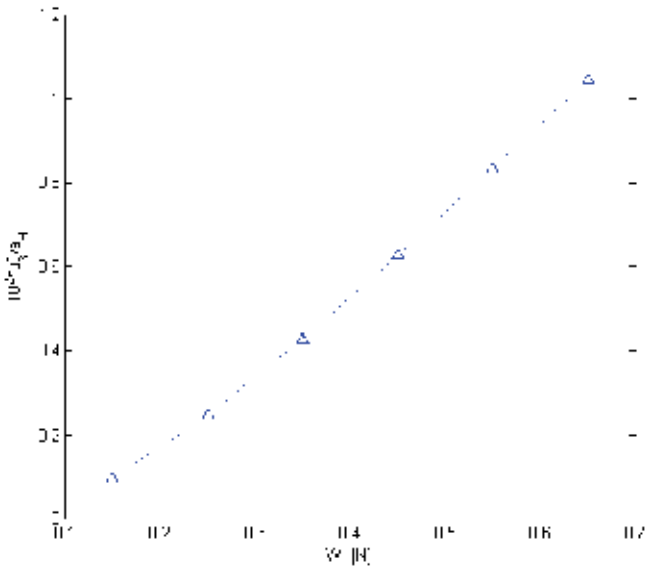


Fig. 11. Residual print depth versus loading level

6.2 E-EP and EP-EP Contact

Normal residual displacement enters interference equation, by superimposing the deflections induced by contact pressure. When only one of the contacting bodies, let it be body (2), is elastic-plastic and the other one, let it be body (1), is elastic, the following interference equation can be written by superimposing the residual part of displacement $u_3^{r(2)}$, related to development of plastic zone in the elastic-plastic body (2), in elastic contact interference relation, Eq. (7):

$$h(i,j) = hi^{(1+2)}(i,j) + u_3^{pr(1+2)}(i,j) + u_3^{r(2)}(i,j) - \omega. \tag{35}$$

On the other hand, when contacting bodies are both elastic-plastic, Eq. (35) encloses residual displacements of both surfaces, namely $u_3^{r(1+2)}(i,j)$. If the hardening behavior or contacting bodies is dissimilar, residual displacement should be computed for every body separately. The model is simplified considerably if the bodies follow the same hardening law and have the same initial contact geometry, because, due to symmetry of the problem about the common plane of contact, $u_3^{r(1)} = u_3^{r(2)}$. Consequently, Eq. (35) becomes:

$$h(i,j) = hi^{(1+2)}(i,j) + u_3^{pr(1+2)}(i,j) + 2u_3^{r(2)}(i,j) - \omega. \tag{36}$$

To validate Eq. (36), the contact between two spheres of radius $R = 0.015m$ is simulated numerically, for two different material behaviors: elastic, and elastic-plastic following Swift's law, with the following parameters: $B = 945MPa$, $C = 20$, $n = 0.121$.
The contact is loaded up to a level of $W = 11{,}179N$, corresponding to a hertzian pressure $p_H = 8GPa$ and to a Hertz contact radius $a_H = 817\mu m$.
Pressure distributions obtained using Eqs. (35) and (36) respectively, depicted in Fig. 12, agree well with already published results, (Boucly et al., 2007). As expected, in the EP-EP contact, pressure appears more flattened compared to the E-EP case, due to a more pronounced increasing in contact conformity related to doubling of the residual term.

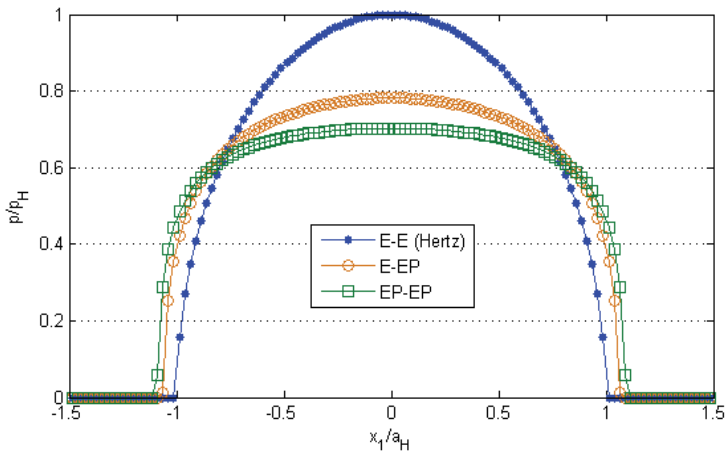


Fig. 12. Pressure profiles for various material behaviors

Variations of maximum effective plastic strain with loading level, in the E-EP and in the EP-EP contact respectively, are depicted in Fig. 13. Intensity of plastic strains in the E-EP contact is up to 40% higher than the one corresponding to the EP-EP scenario.
Variations of maximum pressure with the loading level in the E-E, the E-EP and the EP-EP contact, are depicted in Fig. 14. The curves presented in Figs. 13 and 14 also match well the results of Boucly, Nélias, and Green, (Boucly et al., 2007).

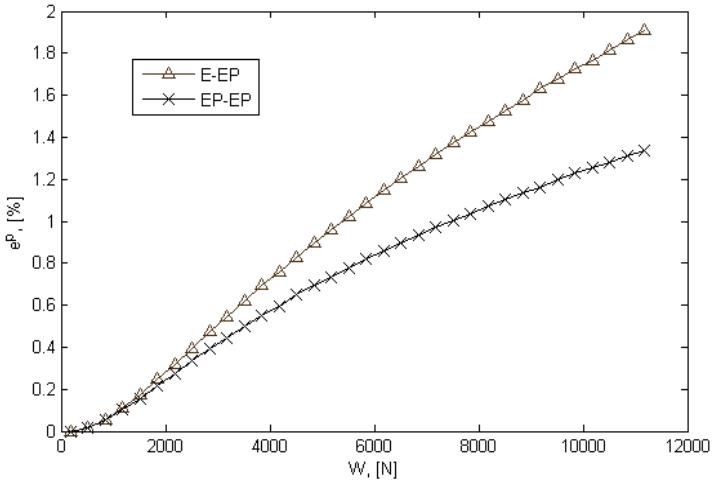


Fig. 13. Maximum effective accumulated plastic strain versus loading level

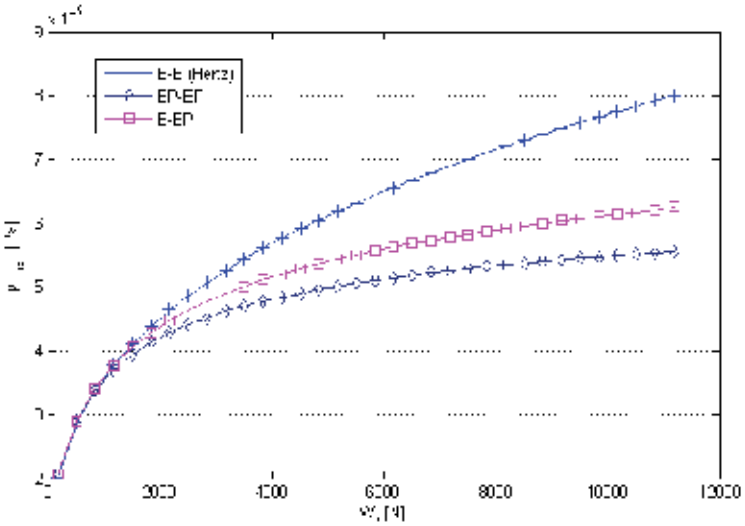


Fig. 14. Maximum pressure versus loading level

6.3 R-EPP contact and experimental validation

As Contact Mechanics uses simplifying assumptions in order to circumvent the mathematical complexity of the arising equations, experimental validation is needed to verify model viability. An extended program of experimental research was conducted in the Contact Mechanics Laboratory of the University of Suceava, aiming to assess residual print parameters in rough elastic-plastic non-conforming contacts. The stand used for the loading experiments was originally designed by Nestor et al., (Nestor et al., 1996). Microtopography of deformed surface was scanned with a laser profilometer UBM14.

Contact between a steel ball, assumed as a rigid indenter, and a lead specimen, simulating the elastic-plastic half-space, was loaded up to an equivalent hertzian pressure

$p_H = 0.94\ GPa$. The contact was also simulated using the numerical formulation. As lead is best described as an EPP material, a linear hardening law with a very small slope was considered in the numerical model. As stated in (Jacq, 2001), the plastic strain increment is undefined when assuming a purely EPP material behavior.
Residual prints at a hertzian pressure of $0.94\ GPa$ is depicted in Fig. 15.

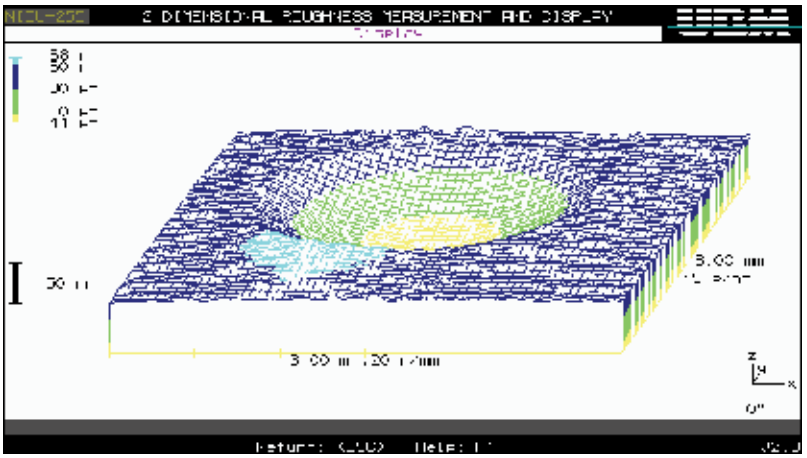


Fig. 15. Experimental residual print in R-EPP contact, $p_H = 0.94\ GPa$

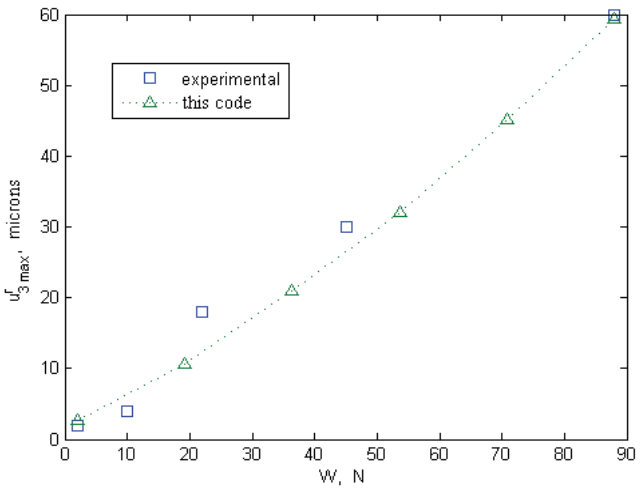


Fig. 16. Residual print depth versus loading level

Variation of print depth with loading level is presented in Fig. 16. The agreement between the values predicted numerically and those obtained experimentally is considered satisfactory, giving the complexity of the phenomena involved.

6. Conclusions

A numerical approach for simulating the elastic-plastic contact with isotropic hardening, based on Betti's reciprocal theorem, is overviewed in this paper. Problem decomposition, as originally suggested by Mayeur and later by Jacq, is employed to assess pressure and plastic strain distribution, on three nested iterative levels.

The newly proposed algorithm has two major advantages over other existing methods. Firstly, the plastic strain increment is determined in a fast convergent Newton-Raphson procedure which iterates a scalar, namely the effective accumulated plastic strain. The method, originally suggested by Fotiu and Nemat-Nasser, employs an elastic predictor, which places the trial state outside yield surface, and a plastic corrector, used to derive the return path to the yield locus. The algorithm is fast, stable, and accurate even for large loading increments.

An additional advantage arises from moving residual stress computation, which is very computationally intensive, to an upper iterative level.

Secondly, the use of three-dimensional spectral methods for solving the intrinsically three-dimensional inclusion problem improves dramatically the overall algorithm efficiency. Solution is obtained by problem decomposition, following a method originally suggested by Chiu. Subproblem of stresses due to eigenstrains in infinite space is solved using influence coefficients also derived by Chiu. Traction-free surface condition is imposed with the aid of Boussinesq fundamental solutions, in a simplified formulation, well adapted to elastic-plastic contact modeling.

With the newly advanced three-dimensional convolution and convolution-correlation hybrid algorithm, based on the DCFFT technique, the computational effort is reduced dramatically, allowing for finer grids in problem discretization.

The newly proposed algorithm was used to simulate, with a high resolution of $120 \times 120 \times 80$ elementary cells, the spherical contact between bodies with various behaviors: R-EP, E-EP, EP-EP and R-EPP.

Elastic-plastic pressure appears flattened compared to the elastic case, due to changes in hardening state of the EP material, and in contact conformity due to superposition of residual displacement in interference equation.

Plastic zone, initially occupying a hemispherical region located at hertzian depths, advances toward half-space boundary with increased loading, enveloping an elastic core. This development is consistent with existing models for the elastic-plastic process, marking the passing from elastic-plastic range to fully plastic.

Residual stress intensity is one order of magnitude smaller than equivalent stresses induced by contact pressure. They contribute to total elastic field by decreasing the peaks in contact stress intensity, thus impeding further plastic flow.

A modified interference equation is used for solving the EP-EP contact with similar hardening behavior and symmetry about the common plane of contact.

Furthermore, residual displacement predicted numerically for the R-EPP contact match well print depths obtained experimentally in indentation of a lead specimen, assumed as an EPP half-space, with a steel ball assumed as a rigid indenter.

7. Acknowledgement

This paper was supported by the project "Progress and development through post-doctoral research and innovation in engineering and applied sciences – PRiDE - Contract no. POSDRU/89/1.5/S/57083", project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.
Grant CNCSIS 757/2007-2008, entitled "Research upon the Effects of Initial Stresses in Dental Biocontacts" provided partial support for this work.

8. References

Antaluca, E. (2005). Contribution a l'étude des contacts élasto-plastiques - effet d'un chargement normal et tangentiel. Ph.D. Thesis, INSA Lyon, France.

Benchea, M. & Cretu, S. (2008). An Improved Incremental Model to Analyse Elastic - Plastic Concentrated Contacts – The Finite Element Analysis and Validation. *Acta Tribologica*, Vol. 16, ISSN 1220-8434.

Boucly, V., Nélias, D., & Green, I. (2007). Modeling of the Rolling and Sliding Contact Between Two Asperities. *J. Tribol. (Trans. ASME)*, Vol. 129, pp. 235 - 245.

Boussinesq, J. (1969). *Application des potentiels á l'etude de l'equilibre et du mouvement des solides élastiques*. Reed. A. Blanchard, Paris.

Chen, W. W., Wang, Q. J., Wang, F., Keer, L. M., & Cao, J. (2008). Three-Dimensional Repeated Elasto-Plastic Point Contacts, Rolling, and Sliding. *J. Tribol. (Trans. ASME)*, Vol. 75, pp. 021021-1 - 021021-12.

Chiu, Y. P. (1977). On the Stress Field Due to Initial Strains in a Cuboid Surrounded by an Infinite Elastic Space. *J. Appl. Mech. (Trans. ASME)*, Vol. 44, p. 587-590.

Chiu, Y. P. (1978). On the Stress Field and Surface Deformation in a Half Space with Cuboidal Zone in Which Initial Strains Are Uniform. *J. Appl. Mech. (Trans. ASME)*, Vol. 45, p. 302-306.

El Ghazal, H. (1999). Etude des proprietes microstructurales et mecaniques des aciers 16NiCrMo13 cemente et 32CrMoV13 nitrure - Application a la prevision de leur limite d'endurance en fatigue de roulement. Ph.D. Thesis, INSA Lyon, France.

Fotiu, P. A., & Nemat-Nasser, S. (1996). A Universal Integration Algorithm for Rate-Dependant Elastoplasticity. *Comput. Struct.*, Vol. 59, pp. 1173–1184.

Jacq, C. (2001). Limite d'endurance et durée de vie en fatigue de roulement du 32CrMoV13 nitruré en présence d'indentations. Ph.D. Thesis, INSA Lyon, France.

Jacq, C., Nélias, D., Lormand, G., & Girodin, D. (2002). Development of a Three-Dimensional Semi-Analytical Elastic-Plastic Contact Code. *J. Tribol. (Trans. ASME)*, Vol. 124, pp. 653–667.

Jin, X., Keer, L. M., and Wang, Q. (2008). Note on the FFT Based Computational Code and Its Application. *Proceedings of the STLE/ASME International Joint Tribology Conference IJTC2008*, October 20-22, 2008, Miami, Florida, USA.

Kalker, J. J., van Randen, Y. A.. (1972). A Minimum Principle for Frictionless Elastic Contact with Application to Non-Hertzian Half-Space Contact Problems. *J. Eng. Math.*, Vol. 6(2), pp. 193-206.

Liu, S. B., and Wang, Q. (2002). Studying Contact Stress Fields Caused by Surface Tractions With a Discrete Convolution and Fast Fourier Transform Algorithm. *J. Tribol. (Trans. ASME)*, Vol. 124, pp. 36–45.

Liu, S. B., Wang, Q., & Liu, G. (2000). A Versatile Method of Discrete Convolution and FFT (DC-FFT) for Contact Analyses. *Wear*, Vol. 243 (1–2), pp. 101–111.

Liu, S. Wang, Q. (2005). Elastic Fields due to Eigenstrains in a Half-Space. *J. Appl. Mech. (Trans. ASME)*, Vol. 72, p. 871-878.

Mayeur, C. (1995). Modélisation du contact rugueux élastoplastique. Ph.D. Thesis, INSA Lyon, France.

Mindlin, R. D., & Cheng, D. H. (1950). Thermoelastic Stress in the Semi-Infinite Solid. *J. Appl. Phys.*, Vol. 21, p. 931-933.

Mura, T. (1968). *The Continuum Theory of Dislocation*. Advances in Material Research, Ed. Herman, H., Vol. 3, Interscience Publisher.

Mura, T. (1988). Inclusion Problem. *ASME Applied Mechanics Review*, Vol. 41, pp. 15-20.

Nélias, D., Boucly, V., & Brunet, M. (2006). Elastic-Plastic Contact Between Rough Surfaces: Proposal for a Wear or Running-In Model. *J. Tribol. (Trans. ASME)*, Vol. 128, pp. 236 - 244.

Nestor, T., Prodan, D., Pătraş-Ciceu, S., Alaci, S., & Pintilie, D. (1996). Stand pentru determinarea histerezisului static la solicitarea de contact (in Romanian). *Proceedings of VAREHD 8*, Suceava.

Polonsky, I. A., & Keer, L. M. (1999). A Numerical Method for Solving Rough Contact Problems Based on the Multi-Level Multi-Summation and Conjugate Gradient Techniques. *Wear*, Vol. 231(2), pp. 206–219.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (1992). *Numerical Recipes in C – The Art of Scientific Computing* – Second Edition. Cambridge University Press.

Shewchuk, J. R. (1994). An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. School of Computer Science, Carnegie Mellon University.

Spinu, S. (2008). A Refined Numerical Method for Elastic Contact Problem with a Tilting Torque on the Contact Area. *Acta Tribologica*, Vol. 16, ISSN 1220-8434.

Spinu, S. (2009). Contributions to the Solution of the Elastic-Plastic Normal Contact Problem (in Romanian), Ph.D. Thesis, University of Suceava, Romania.

Spinu, S., Diaconescu, E. (2008). Numerical Simulation of Elastic Conforming Contacts under Eccentric Loading. *Proceedings of the STLE/ASME International Joint Tribology Conference IJTC2008*, Miami, Florida, USA.

Spinu, S., Diaconescu, E. (2009). A Fast Numerical Method to Predict Elastic Fields Due to Eigenstrains in an Isotropic Half-Space - Part I. Algorithm Overview. *The Annals of University "Dunărea de Jos" of Galati*, Fascicle VIII, 2009 (XV), ISSN 1221-4590, Issue 2, Tribology, pp. 191-196.

Spinu, S., Gradinaru, D. & Marchitan, M. (2007). Improvement of Pressure Distribution in Elastic Non-Hertzian Contacts – Numerical Simulations. *Acta Tribologica*, Vol. 15, ISSN 1220-8434.

Wang, F., & Keer, L. M. (2005). Numerical Simulation for Three Dimensional Elastic-Plastic Contact With Hardening Behavior. *J. Tribol. (Trans. ASME)*, 127, pp. 494–502.

Zhou, K., Chen, W. W., Keer, L. M., & Wang, Q. J. (2009). A Fast Method for Solving Three-Dimensional Arbitrarily Shaped Inclusions in a Half-Space. *Comput. Methods Appl. Mech. Engrg.*, Vol. 198, p. 885-892.

12

Simulating the Response of Structures to Impulse Loadings

Soprano Alessandro and Caputo Francesco
Second University of Naples
Italy

1. Introduction

The need to cope with the new problems which are coupled with progress and its challenges has been causing new design and analysis methodologies to appear and develop; thus, beside the original concept of a structure subjected to statically applied loads, new criteria have been devised and new scenarios analyzed. From fatigue to fracture, vibrations, acoustic, thermomechanics, to remember just a few, many new aspects have been studied in course of the years, all taking place in connection with the appearance of new technical or technological problems, or even with the growing of the consciousness of the relevance of such aspects as safety, reliability, maintenance, manufacturing costs and so on.
One of the problems which in the recent years has been increasingly considered as a relevant one is that of the behaviour of structures in the case of impact loading; there are many reasons for such a study: for example, the requirement to ensure a never-too-satisfactory degree of safety for the occupants of cars, trains or even aircrafts in impact conditions, preventing any collision with the interiors of the vehicle, is just one case.
Another case to be mentioned is that connected with mechanical manufacturing or assembling, which is often carried out with such an high speed as to induce impulse loadings into the involved members; in such cases the aim is to obtain a sound result, even a 'robust' one, in the sense that the same result is to be made as independent as possible from the conceivable variations of the input variables, which, in turn, can be only defined on a probabilistic basis, due for example to their manufacturing scatter and tolerances.
Two main aspects arise in such problems, the first being that related to the definition of the mechanical properties of the materials; the analysis of members behaviour under impulsive loading, for example, requires in general the knowledge of the characteristic curves of materials in presence of high strain rates, which is not usually included in the standard tests which are carried out, so that new experimental tests have to be devised in order to obtain the required items. But at the same time new material families are generated daily, for which no test history is available; in the case of plastics and foams, for example, the search for a reliable database is often a very hard task, so that the analyst has to become a test driver, designing even the test which is the most efficient to obtain effectively the data he needs.
The second problem is the one related to the complication of the geometry and that is adding on the complexity of the analysis of the load conditions. In such cases it is just natural and obvious to direct the own attention to numerical methods, thanks to the ever-

increasing capabilities of computers and commercial codes, and, first of all, to Finite Element Methods (FEM).
FEM, as everything else, is no longer what it used to be in the '70s, when it could scarcely afford to deal with rather easy problems in presence of static conditions, at least from a practical point of view and apart from theory. Nowadays there are commercial codes which can deal with some millions of degrees of freedom (dof's) in static as well dynamic load conditions. The development of numerical procedures which, applying lagrangian and eulerian formulations for finite strains and stresses, allow the analysis of non-linear continua, the use of particular routines for time integration and the progress of the theory of constitutive law for new materials are just a few of the elements, which not only let today researchers investigate rare and particular behaviours of structures, but also allow the birth of rather easy-to-use codes which are increasingly adopted in industrial environments.
Even with such capabilities, the use of the classical "implicit finite element method" encounters many difficulties; therefore, one has to use other tools, and first of all the "explicit FEM", which is well fitted to study dynamic events which take place in very short time intervals. That doesn't mean that analysts don't find relevant difficulties when studying the behaviour of structures subjected to impulsive loads; for example, one has usually to use very short steps in time integration, which causes such analyses to be very time-consuming, even more as one has to overcome serious problems in the treatment of the interface elements used to simulate contact and to represent external loads; at last, only first-order elements (four-node quadrilaterals, eight-node bricks, etc.) are available in the present versions of the most popular commercial codes, what requires very fine meshes to model the largest part of members and that in turn asks for even shorter time steps.
In the following sections, after briefly recalling the main aspects of explicit FEM, we illustrate some of the problems encountered in the study of relevant cases pertaining to the fields of metalformig and manufacturing as well as crashworthiness and biomechanical behaviour, all coming from the direct experience of the authors.

2. Main aspects of explicit FEM

Finite element equations can be written according to Lagrangian or Eulerian formulations; in the former the material is fixed to the finite element mesh which deforms and moves with the material; in Eulerian space the finite element mesh is stationary and the "material flows" through this mesh, what is well suited for fluid dynamic problems. As most structural analysis problems are expressed in Lagrangian space, most commercial codes develop their finite element formulation in that space, even if all of them include algorithms based on Arbitrary Lagrangian-Eulerian (ALE) formulation to face fluid-like material simulation.
To solve a problem of a three-dimensional body located in a Lagrangian space, subjected to external body forces $b_i(t)$ (per unit volume) acting on its whole volume V, traction forces $t_i(t)$ (per unit area) on a portion of its outer surface S_t, and prescribed displacements $d_i(t)$ on the surface S_d, one must seek a solution to the equilibrium equation:

$$\sigma_{ij,j} + \rho b_i - \rho \ddot{x}_i = 0 \,, \tag{1}$$

satisfying the traction boundary conditions over the surface S_t:

$$\sigma_{ij} n_j = t_i(t) \,, \tag{2}$$

and the displacement boundary conditions over S_d:

$$x_i\left(X_a,t\right)=d_i\left(t\right), \tag{3}$$

where σ_{ij} is Cauchy's stress tensor, ρ is the material density, n_j is the outward normal unit vector to the traction surface S_t, X_α (α=1,2,3) and x are the initial and current particle coordinates and *t* is current time.

These equations state the problem in the so-called "strong form", which means that they are to be satisfied at every point in the body or on its surface; to solve a problem numerically by the finite element method, however, it is much more convenient to express equilibrium conditions in the "weak form" where the conditions have to be met only in an average or integral sense.

In the weak form equation, we introduce an arbitrary virtual displacement δx_i that satisfies the displacement boundary condition in S_d. Multiplying equilibrium equation (1) by the virtual displacement and integrating over the volume of the body yields:

$$\int_V\left(\sigma_{ij,j}+\rho b_i-\rho\ddot{x}_i\right)\delta x_i dV=0, \tag{4}$$

by operating simple substitutions and applying traction boundary condition, eq. (4) can be reworked as:

$$\int_V\rho\ddot{x}_i\delta x_i\,dV+\int_V\sigma_{ij}\delta x_{i,j}\,dV-\int_V\rho b_i\delta x_i\,dV-\int_{S_t}t_i\delta x_i\,dS=0 \tag{5}$$

which represents the statement of the principle of virtual work for a general three-dimensional problem.

The next step in deriving the finite element equations is spatial discretization. This is achieved by superimposing a mesh of finite elements interconnected at nodal points. Then shape functions (N_α) are introduced to establish a relationship between the displacements at inner points of the elements and those at the nodal points:

$$\delta x_i=\sum_{a=1}^{n}N_a\delta x_{ai} \tag{6}$$

This task governs all numerical formulations based on the finite element method, whose equations are obtained by discretizing the virtual work equation (5) and replacing the virtual displacement with eq. (6) between the displacements at inner points in the elements and the displacements at the nodal points:

$$\sum_{m=1}^{M}\left\{\int_{V_m}\rho N_a N_\beta\,dV_m\right\}\ddot{x}_{\beta i}=\sum_{m=1}^{M}\int_{V_m}N_a\rho b_i\,dV_m+\sum_{m=1}^{M}\int_{S_t}N_a t_i\,dS_m-\sum_{m=1}^{M}\int_{V_m}N_{a,j}\sigma_{ij}\,dV_m \tag{7}$$

where M is the total number of elements in the system and V_m is the volume of an element.

In matrix form, eq. (7) becomes:

$$\left[\mathbf{M}\right]\left\{\ddot{\mathbf{x}}\right\}=\left\{\mathbf{F}\right\} \tag{8}$$

where [**M**] is the mass matrix, $\ddot{\mathbf{x}}$ is the acceleration vector, and {**F**} is the vector summation of all the internal and external forces. This is the finite element equation that is to be solved at each time step.
The time interval between two successive instants, t_{n-1} and t_n , is the time step $\Delta t_n = t_n - t_{n-1}$; in numerical analysis, integration methods over time are classified according to the structure of the time difference equation. The difference formula is called explicit if the equation for the function at time step *n* only involves the derivatives at previous time steps; otherwise it is called implicit. Explicit integration methods generally lead to solution schemes which do not require the solution of a coupled system of equations, provided that the consistent mass matrix is superseded by a lumped mass one, which offers the great advantage to avoid solving any system equations when updating the nodal accelerations.
In computational mechanics and physics, the central difference method is a popular explicit method.
The explicit method, however, is only conditionally stable, i.e. for the solution to be stable, the time step has to be so small that information do not propagate across more than one element per time step. A typical time step for explicit solutions is in the order of 10^{-6} seconds, but it is not unusual to use even shorter steps. This restriction makes the explicit method inadequate for long dynamic problems. The advantages of the explicit method are that the time integration is easy to implement, the material non-linearity can be cheaply and accurately treated, and the computer resources required are small even for large problems. These advantages make the explicit method ideal for short-duration nonlinear dynamic problems, such as impact and penetration.
The time step of an explicit analysis is determined as the shortest stable time step in any deformable finite element in the mesh. The choice of the time step is a critical one, since a large time step can result in an unstable solution, while a small one can make the computation inefficient: therefore, an accurate estimation has to be carried out.
Generally, time steps change with the current time; this is necessary in most practical calculations since the stable one will change as the mesh deforms. This aspect can make the total runtime unpredictable, even if some "tuning algorithms" implemented in the most popular commercial codes try to avoid it; for example, as that change is required if high deformations are very localized in the model, one can add some masses to the nodes in the deformed area, but not so much to influence the global dynamic behaviour of the structure.
The same tuning process, which leads to added mass to the initial model in those areas where the element size is smaller, can be used to allow an initial time step which is longer than the auto-calculated one. As stated above, the critical time step has to be small enough such that the stress wave does not travel across more than one element at each time step.
This is achieved by using the Courant criteria:

$$\Delta t_e = l/c \tag{9}$$

where Δt_e is the auto-calculated critical time step of an element in the model, l is the characteristic length, and c is the wave speed. The wave speed, c, can be expressed as:

$$c = \sqrt{\frac{E}{\rho\left(1-\nu^2\right)}} \tag{10}$$

where E, ρ and ν are the Young's modulus, density and Poisson's ratio of the material respectively. Therefore, increasing ρ results in an artificial decrease of c and in a parallel increase of Δt_e, without varying the mechanical properties of the material.
The time step of the system is determined by taking the minimum value over all elements:

$$\Delta t_{n+1} = a \cdot \min\left\{\Delta t_1, \Delta t_2, \Delta t_3, \ldots, \Delta t_M\right\} \tag{11}$$

where M is the number of elements. For stability reasons, the scale factor α is typically set to a value of 0.9 (the default in the most popular commercial code, as for example in the LS-Dyna® code) or some smaller value.
Another aspect to be strongly considered when we deal with explicit finite element method is the contact definition, which allows to model the interactions between one or more parts in a numerical model and which is needed in any large deformation problem. The main objective of the contact interfaces is to eliminate any `overlap` or `penetration` between the interacting surfaces. Depending on the type of algorithm used to remove the penetration, both energy and momentum are preserved.
The contact algorithms can be mainly classified into two main branches, one using the penalty methods, which allow penetration to occur but penalize it by applying surface contact force models; the other uses the Lagrange multiplier methods which exactly preserve the non-inter-penetration constraint.
The penalty approach satisfies contact conditions by first detecting the amount of penetration and then applying a force to remove them approximately; the accuracy of approximate solutions depends strongly on the penalty parameter, which is a kind of "stiffness" by which contact surfaces react to the reciprocal penetration. This method is widely used in complex three-dimensional contact–impact problems since it is simple to use in a finite-element solving system. However, there are no clear rules to choose the penalty parameter, as it depends on the particular problem considered. On the other hand, the penalty method affects the stability of the explicit analysis, which is only conditionally stable, when the penalty parameter reaches a certain value with reference to the real stiffness of the material of the interacting surfaces.
Unlike the penalty method, the Lagrange multiplier method doesn't use any algorithmic parameters and it enforces the zero-penetration condition exactly. Thus, this method can give out very accurate displacement fields in the analysis of static contact problems; however, for dynamic contact problems it requires the solution of implicit augmented systems of equations, which can become computationally very expensive for large problems and therefore it is rarely used in solid mechanics field.
Effectively, a contact is defined by identifying what locations are to be checked for potential penetration of a slave node through a master segment. A search for penetrations, using the chosen algorithm, is made every time step. In the case of a penalty-based contact, when a penetration is found a force proportional to the penetration depth is applied to resist, and ultimately to eliminate, the penetration. Rigid bodies may be included in any penalty-based contact but if contact force are to be realistically distributed, it is recommended that the mesh defining any rigid body are as fine as those of any deformable body.
Though sometimes it is convenient and effective to define a single contact to handle any potential contact situation in a model, it is admissible to define a number whatever of contacts in a single model. It is generally recommended that redundant contacts, i.e., two or more contacts producing forces due to the same penetration (for example near a corner), are

avoided, as this can lead to numerical instabilities. To enable flexibility for the user in modelling contact, commercial codes present a number of contact types and a number of parameters that control various aspects of the contact treatment. But, as already stated, unfortunately, there are no clear rules to choose these parameters, depending from user's experience and, in any case, their values are often obtained by means of trials and error iterative procedure.
Anyway, the best way to start a contact analysis by using a commercial explicit solver is to consider default settings for these parameters, even if often non-default values are more appropriate, to define the same element characteristic lengths to model interacting surfaces and, overall, to avoid initial geometrical co-penetrations of contact surfaces.
Thus, the selection of integration time step and of the contact parameters are two important aspects to be considered when analysts deal with simulation of the response of structure to impulse loading.
The last important topic examined in the present section and which can result in additional CPU costs as compared to a run where default parameters values are used, regards shell elements formulation. The most widely adopted shells in commercial codes belong to the families of the Hughes-Liu or of the Belytschko-Tsay shell elements. The second one is computationally more efficient due to some mathematical simplifications (based on co-rotational and velocity-strain formulations), but results in some restriction in the computation of out of plane deformations.
But the real problem is that, in order to further reduce CPU time, analysts generally aims to use under integrated shell elements (i.e. with a single integration point), and this causes another numerical problem, which also arises with under-integrated solid elements. This numerical problem concerns the hourglassing energy: single integration point elements can shear without introducing any energy, therefore an added "numerical energy" is generated to take it into account. High hourglassing energy is often a sign that mesh issues may need to be addressed by reducing element size, but the only way to entirely eliminate it is to switch to formulations with fully-integrated or selectively reduced integration (S/R) elements; unfortunately, this approach is much more time expensive and can be unstable in very large deformation applications, therefore hourglassing energy is generally controlled by considering very regular meshes or by considering some corrective algorithms provided by commercial explicit solvers. In any case, these algorithms ask for an analysts much experienced on their formulation, otherwise other numerical instabilities can arise following their use.

3. Some case studies from manufacturing

Some case studies are now presented to introduce the capabilities and peculiarities of the analysis of structures subjected to impulsive loadings; they are connected with some of the relevant problems of manufacturing and will let the reader to grasp the basic difficulties encountered, for example, when dealing with contact elements which model interfaces. The first one deals with the case of riveted joints and shows how to simulate the riveting operation and its influence on the subsequent bulging coming from an axial load, while the second one comes from metalforming and deals with the stretch-bending process of an aluminium C-shaped beam.

3.1 The analysis of the riveting process

The load transfer mechanism of joints equipped with fasteners has been recognized for a long time as one of the main causes which affect both static resistance as well as fatigue life of joints; unfortunately, such components, which are often considered as very simple, exhibit such a complex behaviour that it is far from being deeply understood and only in recent times the coupling of experimental tests with numerical procedures has let researchers begin to obtain some knowledge about the effects which come from assuming one of the available designs.

Starting from the very simple hypothesis about load transfer mechanisms which are used in the most common and easy cases, a real study of such joints has started just after Second World War, mainly because from those years onward the use of bolted or riveted sheets has been increasingly spreading and several formulae were developed with various means; also in those years the "neutral line method" was introduced to study the behaviour of the whole joint, with the consequence that the need of a sound evaluation of fasteners stiffness and contribution to the overall behaviour was strictly required. A wide spectrum of results and theories have appeared since then, each one with some peculiarities of its own and the analysis of bolted and riveted joints appears now as to be analysed by different methods.

The requirement of a wide range of different studies is to be found in the large number of variables which can affect the response of such joints, among which we can quote, from a general but not exhaustive standpoint:

- ***general parameters***: geometry of the joint (single or several rows, simple- or double-lap joints, clamping length, fastener geometry); characteristics of the sheets (metallic, non metallic, degree of anisotropy, composition of laminae and stacking order for laminates); friction between sheets, interlaminar resistance between laminae, possible presence of adhesive;
- ***parameters for bolted joints:*** geometry of heads and washers; assembly axial load; effective contact area between bolts and holes; fit of bolts in holes;
- ***parameters for riveted joints:*** geometry of head and kind of fastener (solid, blind – or cherry – and self-piercing rivets, besides the many types now available); amplitude of clearance before assembly; mounting axial load; pressure effects after manufacture.

From all above it follows that today a great interest is increasingly being devoted to the problem of load transfer in riveted joints, but that no exhaustive analysis has been carried out insofar: the many papers which deal with such studies, in fact, analyze peculiar aspects of such joints, and little efforts have been directed to the connection between riveting operation and response of the joint, especially with regard to the behaviour in presence of damage.

Therefore, the activity which we are referring to dealt with modelling of the riveting operation, in order to define by numerical methods the influence of the assembly conditions and parameters on the residual stress state and to the effective compression zone between sheets; another aspect to be investigated was the detection of the relevant parameters of the previous operation to be taken into account in the analysis of the joint strength.

As we wished to analyse the riveting operation and its consequences on the residual stresses between plates, the obvious choice was to use a dynamic explicit FEM code, namely Ls-Dyna®, whose capabilities make it most valuable to model high-speed transients without much time consumption.

As a drawback, we know that that code is very sensitive to contact problems and that a finer mesh requires smaller integration time intervals: therefore the building of a good model, parametrically organized in order to make variations of input parameters easy, took a long time. The procedure we followed was to use ANSYS® 10.0 PDL (parametric design language) capabilities to be coupled with Ls-Dyna solver to obtain a global procedure which can be summarized in the following steps:

- Write a parametric input file for ANSYS PDL, where geometry, behaviour of materials, contact surfaces and conditions, load cases were specified; it gives a first approximate and partially filled Ls-Dyna input file;
- Complete the input file for Ls-Dyna, in order to introduce those characteristics and instructions which are required, but which are not present in Ansys code, mostly control cards and some variations on materials;
- Solve the model by Ls-Dyna code;
- Examine the results by Ls-PrePost or by Ansys post-processor module, or by Hyperview® software, according to the particular requirements.

In fig. 1 one can see the basic Ls-Dyna model built for the present analysis, with reference to a solid rivet; the model is composed of seven parts, among which one can count three solid parts, made of brick elements, and four parts composed by shells: three of these are required to represent the contact surface, while the last composes a plane rigid wall that represents the riveting apparatus.

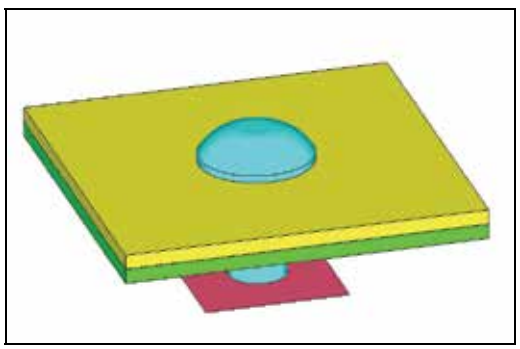


Fig. 1. The model used to simulate the joint

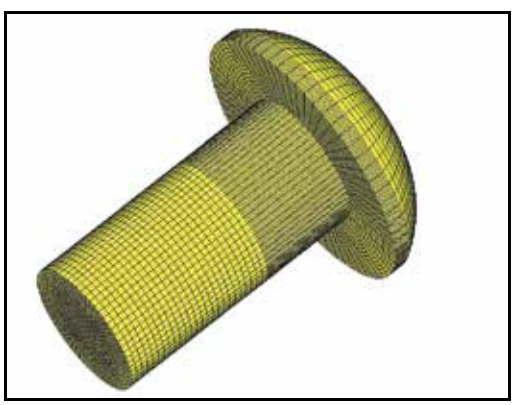


Fig. 2. The model of the rivet

A finer mesh – with a 0.2 mm average length – was adopted to model the stem of the rivet (fig. 2) and those parts of the sheets which, around the hole and below the rivet head, are more interested by high stress gradients; a coarser mesh was then adopted for the other zones, as the rivet head and the parts of the sheets which are relatively far from the rivet.
The whole model was composed, in the basic reference case, of 101,679-109,689 nodes and 92,416-100,096 brick elements, according to the requirements of single cases, which is quite a large number but also in that case runtimes were rather long, as they resulted to be around 9-10 hours on a common desktop; more complex cases were run on a single blade of an available cluster, equipped with 2 Xeon 3.84 GHz - 4 GB RAM - and of course comparatively shorter times were obtained.
The main reason of such times is to be found in the very short time-step to be used for the solution, about 1.0E-08 s, because of the small edge length of the elements.
The solid part of rivet and sheets were modelled following a material 3 from Ls-Dyna library, which is well suited to model isotropic and kinematic hardening plasticity, with the option of including strain rate effects; values were assigned with reference to 2024 aluminium alloy; the shells corresponding to the contact surfaces were then modelled with a material 9, which is the so-called "null material", in order to take into account the fact that those shells are not a part of the structure, but they are only needed to "soften out" contact conditions; for that material shells are completely by-passed in the element stiffness processing, but not in the mass processing, implying an added mass, and for that reason one has to manually assign penalty coefficients in the input file. Some calibration was required to choose the thickness of those elements, looking for a compromise between the influence of added mass – which results from too large a thickness – and the negative effect with regard to contact, which comes in presence of a thickness too small, as in that case Ls-Dyna code doesn't always detect penetration.
The punching part was modelled as a rigid material (mat. no. 20 from Ls-Dyna library); such a material is very cost effective, as they, too, are completely bypassed in element processing and no space is allocated for storing history variables; also, this material is usually adopted when dealing with tooling in a forming process, as the tool stiffness is some order larger than that of the piece under working. In any case, for contact reasons Ls-Dyna code expects to receive material constants, which were assumed to be about ten times those of steel.
For what concerns the size of the rivet, it was assumed to be a 4.0 mm diameter rivet, with a stem at least 8.0 mm long; as required by the general standards, considering the tolerance range, the real diameter can vary between 3.94 and 4.04 mm, while the hole diameter is between 4.02 and 4.11 mm, resulting in diametral clearances ranging from 0.02 to 0.17 mm; three cases were then examined, corresponding to 0.02-0.08-0.17 mm clearances.
The sheets, also made of aluminium alloy, were considered to range from 1.0 to 4.0 mm thickness, given the diameter of the rivet; the extension examined for the sheets was assumed to correspond to a half-pitch of the rivets and, in particular, it was assigned to be 12.5 mm; along the thickness, a variable number of elements could be assigned, but we considered it to be the same of the elements spacing along the stem of the rivet: that was because contact algorithms give the best results if such spacing is the same on the two sides of the contact region. In general, we introduced a 0.2 mm edge length for those elements, which resulted in 5 elements along the thickness, but also case of 10 and 20 elements were investigated, in order to check the convergence of the solution.
At last, for what concerns the loads, they were applied imparting an assigned speed to the rigid wall, and recovering *a posteriori* the resulting load; that was because previous

experiences suggested not to directly apply forces; besides, all applicable loads accepted by Ls-Dyna are body forces, or one concentrated force on a rigid body, or nodal forces or pressure on shell elements: the last two choices don't guarantee the planarity of the loaded end after deformation, which can be obtained by applying the load on the tool, but that use in past experiences revealed to be rather difficult to be calibrated.
Therefore, we assumed a hammer speed-time law characterized by a steep rise in about 0.006 s up to the riveting speed, which remains constant for a convenient time, then subduing an inversion also in about 0.006 s after the wanted distance has been covered; considering that the available data mention 0.2 s as a typical riveting time, the tool speed has been assumed to be 250 mm/s, even if the effects of lower velocities were examined (200, 150 and 50 mm/s).
Therefore, summarizing the analyses carried out insofar, the variables assumed were as follows:

- Initial clearance between the rivet stem and the hole;
- Thickness of the sheets;
- Speed of the tool.

The results obtained can be illustrated, first of all, by means of some countour plots, beginning from fig. 3 and 4, where the variation of von Mises equivalent stress is illustrated for the cases defined above, concerning the clearance amplitude between rivet and hole; it is quite evident, indeed, that the general stress state for the max clearance case is well below what happens when the gap decreases, also considering the scale max values: the mean stress level in sheets increases, as well as the largest absolute values, which can be found in correspondence of the folding of the rivet against the edge of the hole.

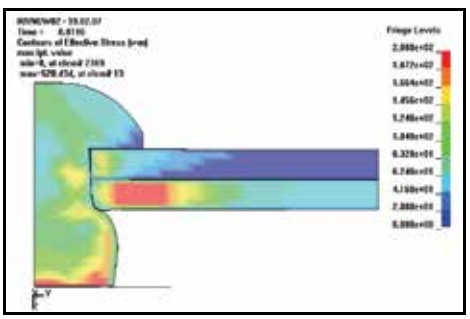


Fig. 3. Von Mises stress during riveting for max clearance

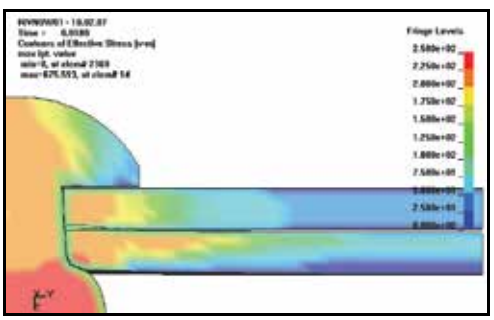


Fig. 4. Von Mises stress during riveting for min clearance

While the previous results have been illustrated with reference to the time when the largest displacement of the rigid wall occurs, others can be best observed considering the final time, when the tool has left the rivet and possible stress recovery determined.
For example, it can be useful to look at the distribution of pressure against the inner surface of the hole for the same cases above. The results observed can be summarized considering that in presence of the max clearance the rivet can fill the hole completely – and that the second sheet is only partially subjected to internal load – and then all the load is absorbed from the first edge of the hole, which is therefore overstressed, as a part of the wall doesn't participate to balance load; also the external area of the first sheet interested by the folding of the rivet is quite large.
When clearance reduces it can be observed that gradually all the internal surface of the hole comes in contact with the rivet and therefore it can exert a stiffening action on the stem, which folds in a lesser degree and therefore can't transmit a very large load on the edge of the hole, as it can be observed in fig. 5 as the volume of the sheet which is subjected to significant radial stresses.

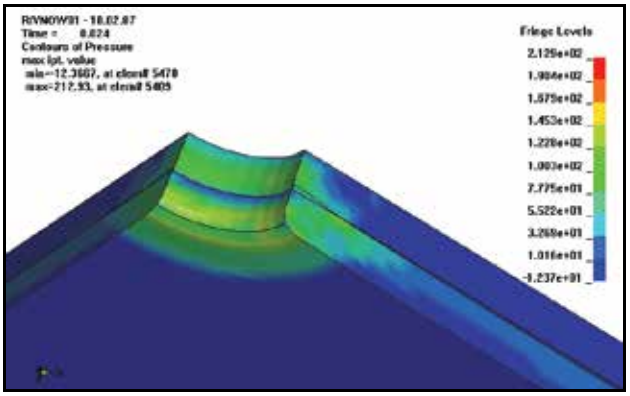


Fig. 5. Residual pressure for min clearance

Also the extension of the volume interested by plasticity increases; in particular we obtained that in presence of a larger gap only a part of the first sheet is plastically deformed, but, at the same time, that the corresponding deformation reaches higher values, all in correspondence of the external edge or immediately near to it; as clearance reduces the max plastic deformation becomes smaller, but plasticity reaches the edge of the second sheet and that effect is still larger in correspondence of the min clearance, where a larger part of the second sheet is plastically deformed; at the same time the largest values of the plastic deformation in correspondence of the first edge becomes moderately higher for the constraint effect exerted by the inner surface of the hole and above noted.
It is interesting to notice that the compression load is no much altered by varying the riveting velocity, as it can be observed from fig. 6 for 1.00 mm thick plates; what is more noteworthy is the large decrease from the peak to the residual load, which is, more or less, the same for all cases.
On the other hand, the increase of thickness produces larger compression loads (fig. 7), as it was to be expected, because of the larger stiffness of the elements. It must be noted, for comparison reasons, that for the plots above the load is the one which acts on the whole rivet and not on the quarter model.

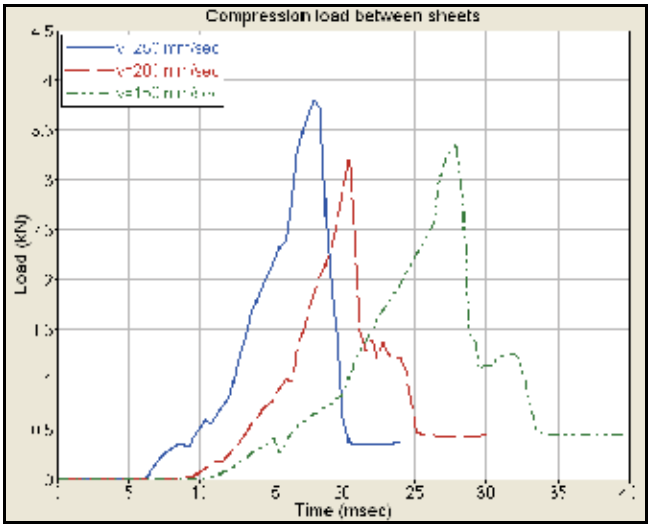


Fig. 6. Influence of velocity on compression load

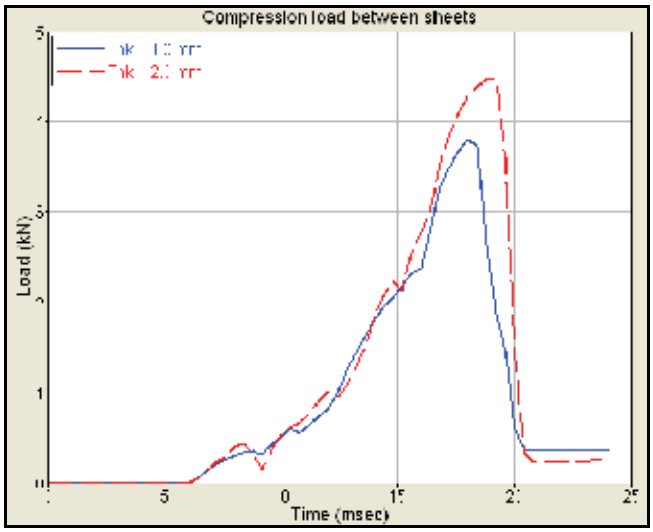


Fig. 7. Influence of thickness on compression load

Aiming to evaluate the consequences of the riveting operation on the behaviour of a general joint, because of the residual stress state which has been induced in the sheets, the effect of an axial load was investigated, considering such high loads as to cause a bulging effect. As a first step, using an apparatus (Zwick Roell Z010-10kN) which was available at the laboratories of the Second University of Naples, a series of bearing experimental tests (ASTM E238-84) have been carried out on a simple aluminium alloy 6xxx T6 holed plate (28.5 x 200 x 3 mm3, hole diam. 6 mm), equipped with a 6 mm steel pin (therefore different from that for which we presented the results in the previous pages) obtaining the response curves shown in Fig. 8. In the same graph numerical results have been illustrated, carried out from non linear static FE simulations developed by using ANSYS® ver. 10 code. As it is

possible to observe the agreement between numerical and experimental results is very good. This experimental activity allowed to setup and develop the FE model (Fig. 9) of each single sheet of the joint and, in particular, their elastic-plastic material behaviour.

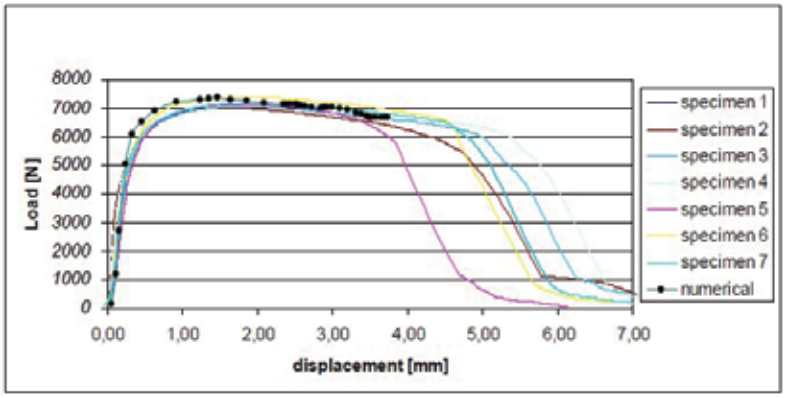


Fig. 8. Results from experimental and numerical bearing tests

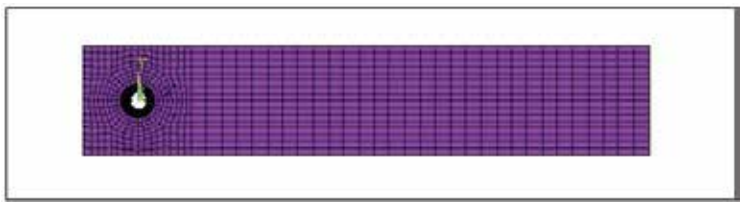


Fig. 9. FE model of a single joint sheet

In order to investigate on the influence of the riveting process, the residual stress-strain distribution around the hole coming from the riveting process above was transferred to the model of the riveted joint (sheets dim. 28.5 x 200 x 1 mm3, hole diam. 6 mm). The transfer procedure consisted in the fitting of the deformed rivet into the undeformed sheets and in the subsequent recovery of the real interference as a first step of an implicit FE analysis.
After the riveting effect has been transferred to the joint the sheets were loaded along the longitudinal direction and the distribution of Von Mises stress around the hole of one sheet of the joint in presence of the maximum value of the axial load value is illustrated in Fig. 10. The results in terms of axial load vs. axial displacement have been compared (Fig. 11) with

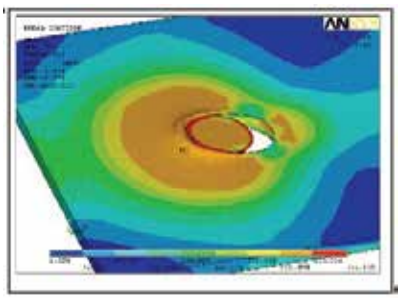


Fig. 10. Bulging of the riveted hole coming from implicit FEM

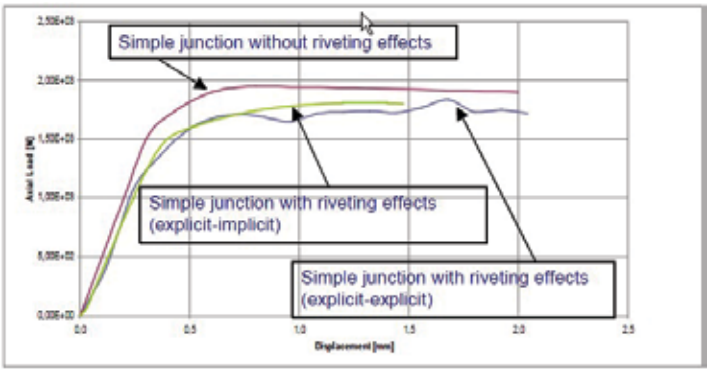


Fig. 11. Effect of the residual stress state on the behaviour of the joint

those previously obtained from the analysis of the same joint without taking into account the riveting effect: it is possible to observe that the riveting operation effects cause a reduction of the bearing resistance of the joint of about 10%. On the same plot also the results obtained by analysing also the axial loading by means of the explicit codes are illustrated: this procedure obviously proved to be very time consuming compared to the use of an explicit to implicit scheme, without giving relevant advantages in terms of results and therefore it is clear that the explicit-implicit formulation can be adopted for such analyses.

3.2 A stretch-bending case study

As it is known, the space frame with the whole load-carrying structure made of aluminium alloy is an assessed concept. A feature of this kind of application is that the originally straight extrusion of some component must be followed by some plastic forming operations in order to obtain the desired shape/curvature. Several types of modified bending processes are thus introduced, e.g. press bending, rotary draw bending, stretch bending, etc.

Typical concerns regarding the industrial use of these methods are the magnitude of the tolerances during production and the cross-sectional distortions of the curved specimen. The tolerance problem is primarily related to the springback phenomenon: springback is the elastic recovery taking place during unloading; the most important cross-sectional distortions are local buckling in the compression zone and sagging, which is a curvature-induced local deformation of the cross-section.

In-house experience combined with trial-and-error procedures has been the traditional solution of the tolerance and distortion challenges in industrial bending. This approach may be time consuming and expensive, therefore alternative methods are requested, including the use of the numerical simulation by means of finite element method.

There are several difficulties associated with a numerical simulation of the stretch bending of extruded components; the main ones are non-linear material behavior, geometrical non-linearities, modeling of boundary conditions, contact between die and specimen, springback during the unloading phase. Another very complex aspect is the calibration of the numerical model as rather few experimental results are available in the literature. In any case, to simulate these typologies of phenomena explicit FE algorithms can be certainly considered the most suitable, for what concerns both the computational efficiency and the solution accuracy; on the other side, implicit FE algorithms can be considered in the most of applications more effective in the spring-back phase.

The experimental test-case regards a process of stretch-bending of a single frame (3000 mm length) of aluminium alloy 7076, whose transversal section is represented in figure 12; during the process the ends of the frame are clamped and a tensile force, corresponding to the yield force or somewhat higher, is applied to the specimen. Then the frame is bended by fitting it around a die (3300 mm radius) with the mandrel fixed and the arms of the machine rotating. Stretch bending of the frame has been developed after it has been subjected to a quenching treatment.

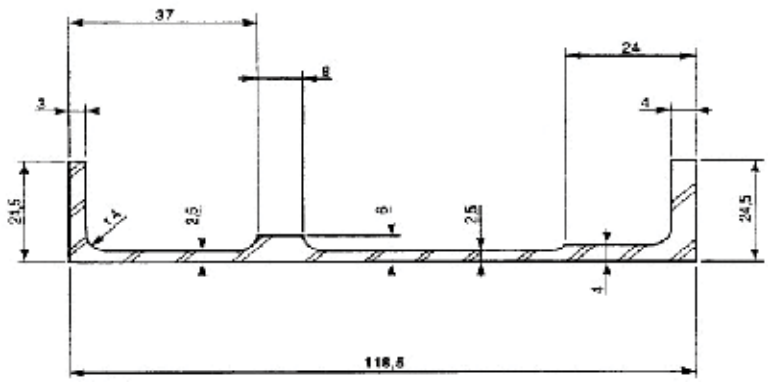


Fig. 12. Transversal section of the bended frame (dimensions are in mm)

In order to evaluate residual stresses after the stretch bending, experimental hole-drilling measurements have been performed in opportune locations on the frame, as showed in figure 13, where also the test apparatus is illustrated.

Fig. 13. Hole drilling measurement locations and test apparatus

The developed FE model consists of 743,000 8-noded hexahedral solid elements (3 dof's per node) and 694,000 nodes. Plastic-kinematic behavior is assumed to model mechanical material properties (E=74000 MPa, v=0.3, σ_y=461 MPa, Etan=700MPa). Only half frame has been modelled because of the symmetry. Some solid elements fit into the frame section have been considered in both the real and the virtual process in order to prevent the sagging deformation due to the buckling of the section.

The stretch bending process has been simulated by considering the mandrel fixed in the space and perfectly rigid. The nodes on the transversal section of the frame belonging to the symmetry plane are constrained to move only in the symmetry plane; the nodes of the end transversal section are rigidly linked to the node of the rigid bar elements representing the arms of the bending apparatus, which rotate and push the frame on the mandrel by following opportune paths. It should be noted that the frame is initially stretched and then

bended. The explicit FE algorithms implemented in the Ls-Dyna® [6] code have been used to develop the loading phase of the analysis.
For what concerns the unloading phase, some attempts have been made to solve it by using implicit algorithms of the Ls-Dyna® code, but a lot of convergence problems have been arisen due to the large relative displacements between the different elements of the chain; to avoid this kind of problems significant model modifications are needed, therefore it has been more convenient to simulate the unloading phase by using explicit finite element algorithms, by introducing a fictitious damping factor. In figure 14, the kinetic energy of the frame vs. process time is showed, where it is possible to individuate the start time of the spring-back phase.

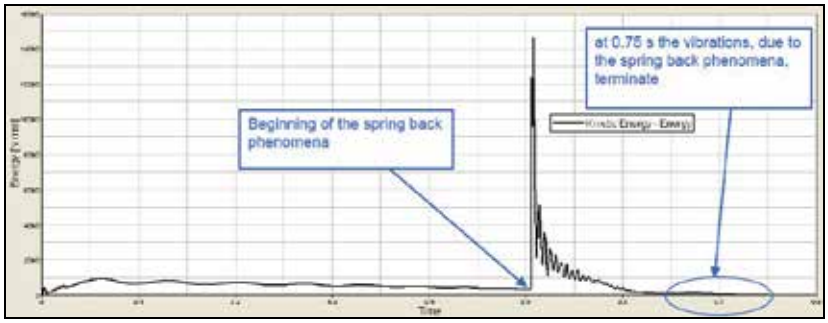


Fig. 14. Kinetic energy of the frame during stretch-bending

4. Biomechanical problems in crashworthiness studies

One of the most relevant issues in today engineering is that related to safety in transportation; as it is a common statement that our lines are as safe and able to avoid any accident in the highest degree, with respect to the actual design and manufacturing procedures, the greatest attention is now being paid to the protection of passengers when unfortunately an impact occurs (i.e. to what is today called the "passive safety").
In those occasions, indeed, passengers can be injured or even killed because of the high decelerations which take place or because they move in the vehicle and impact against the structure or even because the deformation of the structure is so severe as to reduce or even to cancel the required space of survival.
That knowledge has brought designers to introduce sacrificial elements in the structures, i.e. some elements which adsorb the incoming kinetic energy by deformation and slow down decelerations, thus preventing the passengers from severe impacts; in other cases, means restraint such seatbelts are provided, in order to avoid undesired or dangerous motions of the same travellers.
The studies of such dangerous events have shown that the impact occurs in a very short time (typically, 100-150 ms in the case of cars) which explains the large inertia forces which are developed and therefore the analyses have to be carried out in time, i.e. as a transient analysis in presence of finite deformations and of highly non-linear and strain rate dependent materials.
As the aim of such studies is to prevent or at least to limit the damage of passengers, it is obvious that all results are made available in terms of decelerations and impact forces on human bodies, which are to be compared with the respective admissible values, which have

been studied for many years now and are rather well assessed. Specialized centres, as NCAP in the car field, have defined many biomechanical indexes which can now obtained for a given crash scenario from the same numerical analysis of the impact and which can immediately compared with known limit values. For example, the most well known index, HIC (Head Injury Criterion), evaluates the maximum acceleration level which acts for a sufficient time on the neck of a passenger implicated in an accident, according to the following expression:

$$HIC = max\left\{\left[\frac{1}{t_2 - t_1}\int_{t_1}^{t_2} a(t)\cdot dt\right]^{2.5}\cdot(t_2 - t_1)\right\} \tag{12}$$

where a(t) is the total acceleration of the neck which occurs in the interval $t_1 \div t_2$, which usually is assumed to be 36 ms; as that span is shorter than the crash, a window is moved along the time axis up to the point where the largest values of the index are obtained.
Beside HIC, many other indexes have been defined, as VC (Viscous Criterion), TTH (Thorax Trauma Index), TI (Tibia Index) and others, all referring to different parts of the human body; all results are then combined to assess the safety level of the structure (car, train or other) in a particular impact scenario.
The soundness of a structural design which involves safety issues is assessed on that basis and that let us realize the difficulties of the procedure. Beside, one has to realize that the characteristics of the adopted materials have to be precisely known for the particular accident one has to analyze; that means that the behaviour of the materials has to be acquired in the non-linear range, but also in presence of high strain rates. Usually those behaviours are not known in advance and therefore specific tests have to be carried out before the numerical analysis.
At last, because of the simplifying hypotheses one introduces inevitably in the numerical model, it is necessary to calibrate it with reference to some known beforehand particular scenario, to be sure that the behaviour of the material is well modelled.
It has to be stressed that in the past the main way to obtain reliable results was to carry out experimental tests, using anthropomorphic dummies and structures, which suffered such damages as to prevent their further use. That way was very time consuming and implied such unbearable costs that it couldn't be performed on a large scale basis, to examine all possible cases and to repeat test a sufficient number of times; the consequence was that passive safety didn't advance to high standards.
When numerical codes improved to such levels as to manage complex analyses evolving in time in presence of finite stresses and strains, it was only a matter of time before they began to be used to simulate impact scenarios; that has resulted in a better understanding of the corresponding problems and in obtaining a much larger number of results, which in turn allowed an important level of knowledge to be achieved.
Therefore, today activity in passive safety studies is mainly performed by simulation methods and a much lesser number of experimental tests is carried out than in the past. Thus, it is now possible to study very particular and specific cases, but in order to obtain reliable results it is quite necessary to calibrate each analysis with experimental tests and to comply with codes and standards which were often devised when today computers were not yet available.

5. Case studies from crashworthiness analyses

5.1 An example of crashworthiness analysis in the automotive field

In the first development stages of the numerical analysis of vehicle impacts, with studies about the energy absorption capabilities of sacrificial elements and on biomechanical damages, some scenarios were introduced and their understanding deepened, as frontal impact with or without offset, lateral impact, rollover, pole impact and so on.

Those cases are now widely assessed and more particular scenarios are being studied, as that referring to pedestrian impact or that considering the oblique impact against road guardrails; in the present section, however, we introduce a very specific and interesting case, as it can be usefully adopted to clarify the degree of accuracy that is required today.

One of the most interesting cases, indeed, is that referring to the contingency that a passenger, because of his motion in the course of an accident, impacts against one of the fixtures which define the compartment or the many appliances and gadgets which are fitted to its walls or which constitute its structure.

As the most dangerous case is that when it is the passenger's head to be involved in such an impact, the corresponding study is a very relevant one, as one would have to ensure that interior tapestry and its thin foam stuffing, for example, have such energy absorption capabilities as to prevent severe damages to the head when coming into contact with the metal structure of the compartment.

As one of the main advantages of numerical simulation is to reduce the number of physical tests, it is just natural to try and reproduce the experimental conditions and equipments in order to ensure a reliable correlation between the two cases; now, tests are performed on the basis of USA CFR (Code of Federal Regulations), which have been more or less included in EEVC (European Enhanced Vehicle Safety Committee) standards and therefore one has to be sure to comply with them.

The experimental test of such impact is carried out by simply firing a head-shaped impactor against the target in study, hitting it in precisely defined locations along assigned trajectories; such an impactor is just the head of a dummy whose characteristics have to be verified according very strict standards. For example, the head is to be dropped from a height of 376 mm on a rigidly supported flat horizontal steel plate, which is 51 mm thick and 508 mm square; it has to be suspended in such a way as to prevent lateral accelerations larger than 15g to occur and the peak resultant acceleration recorded at the locations of the accelerometers mounted in the headform have to be not less than 225g and not larger than 275g (Fig. 15).

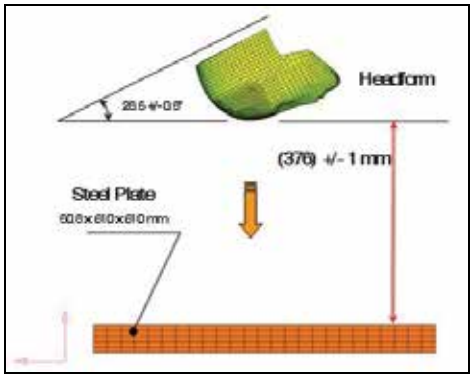


Fig. 15. Headform test conditions

Those standards have hard consequences for the numerical simulations, as one wishes to model the headform as an empty shell-like body, in order to save runtime, but it has to exhibit a stiffness as well as inertia properties such as to be equivalent to the physical head. To respect those conditions and to prevent some wavy dynamic deformations to appear, it can be useful to provide the model with a very stiff ring in the rearmost part (Fig. 16).

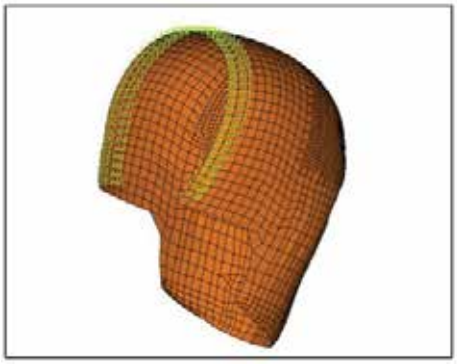


Fig. 16. The stiffened headform

Furthermore, to save time the simulation can start at the time when the impact begins, imparting to the model the same velocity which it would get after falling from the assigned height (Fig. 17). After successfully running the model, an acceleration/time plot is obtained, where the peak values fall in the expected range (Fig. 18).

Once the model of the head has been created and calibrated, one has a large number of difficulties to take into account; beside dashboard, sun visor, header, seat-belt slit, internal handles, there are A- and B-pillars, front header, side rails, and each can be impacted in several points in dependence of the initial position of the passenger. CFR and EEVC show how to define all such points, by means of rules which take into account the geometry of the compartment.

When one comes to a particular obstacle, one has to consider that it is not an easy, single part component; broadly speaking, it is composed by a padding which is mounted on the structure with the interposition of a foam stuffing and the padding usually has several ribs which stiffen the component and position it exactly on the structure. Moreover, the mounting can be obtained by adhesives, clips, rivets or by other means. All that has to be

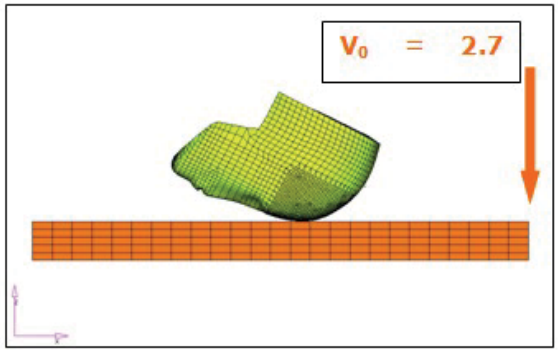


Fig. 17. Imparting an equivalent velocity to the headform

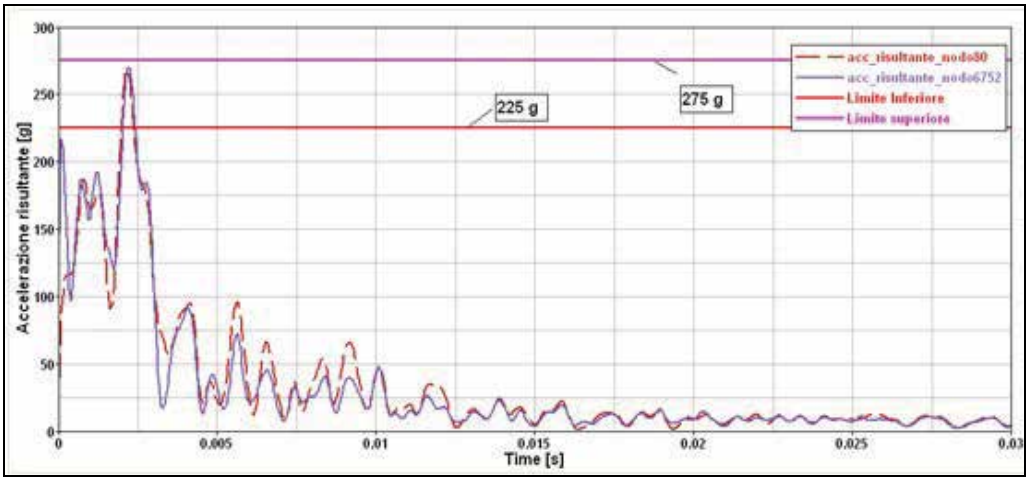


Fig. 18. Resultant acceleration in dropping test

modelled precisely if one wants to get reliable results and beside the modelling elements one has to add the interface ones, which are elements such as to take into account the contact conditions and to prevent copenetrations between the different bodies.
The result is that the modelling task is not at all a secondary one, but it requires a long labour and great attention, also because of the particular shapes which characterize today the various components.
For example, in Fig. 19 it is shown the case of the simulation of the impact of the headform against an upper handle; the use of a code like Ls-Dyna® let the analyst get a complete set of results, such as displacements, velocities and accelerations of each element, as well as contact and inertia forces, beside the energy involved, subdivided in all relevant parts, as kinetic, deformation, and so on.

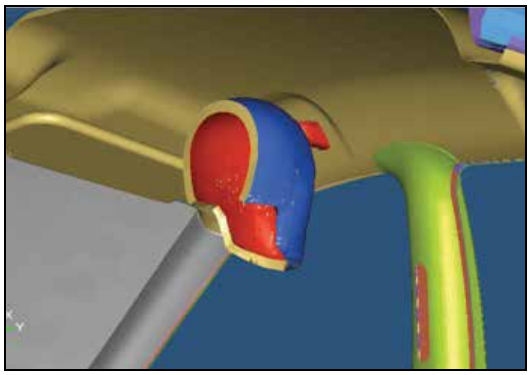


Fig. 19. The impact of the headform against the internal upper handle

Nevertheless, one has to realize that the obtained results are not so smooth as one could guess, because of evaluation and round-off errors, instability of the elements and of the numerical procedure, and so on; once grouped in a plot, the result curves show peaks and valleys which are meaningless and have to be removed, just as one does when dealing with vibration or sound curves; the usual technique is to treat the numerical values with a filter (for example, SAE 100 or 180) which makes the plot more intelligible.

One of the obtained results, for example, is that shown in Fig. 20, which refers to the previous impact case against the handle; four curves are shown, i.e. the experimental one, together with ±15% curves, which bound the admissible errors, and that which comes from numerical simulations; as it can be observed, the numerical values are all inside the admissible range, but for a later time, which comes when the headform has left the obstacle and is moving free in the compartment, which is of no interest.

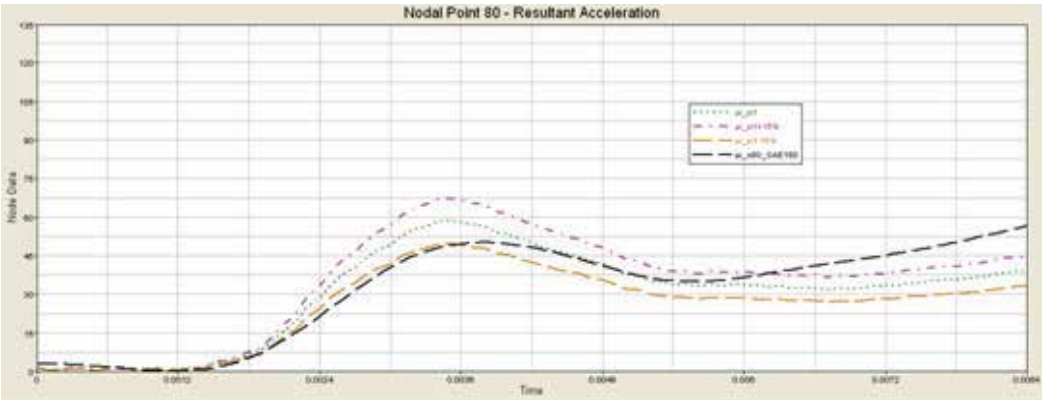


Fig. 20. Resultant acceleration for the impact of the headform against the internal upper handle

5.2 Crashworthiness analyses in railways

The survival of the occupants of a railway vehicle, following an accident, depends substantially from three aspects:

- type and severity of the accident;
- crash behaviour of the structure as a whole;
- resulting type and severity of secondary impacts which occur because of the relative speed between passengers and interiors.

The investigation on these aspects, by means of numerical methods, starts from the identification and the successive simulation of opportune impact scenarios involving a detailed numerical model of the vehicle as a whole. The identification of the most representative impact scenarios is taken from the EN15227 standard (Railway applications - Crashworthiness requirements for railway vehicle bodies). The simulation of the impact scenario provides the evaluation of the deformations suffered from the structure and from the interiors and allows the identification of the kinematic and dynamic properties necessary to set up the biomechanical analyses.

Within this work, the overall resistance of the vehicle was considered fixed, as the mean objective was the simulation of the biomechanical performances of an interior component (hereinafter also called panels), in order to identify its characteristics of passive safety and to assess guidelines to improve its design.

The goal was to set up a hybrid methodology which uses in a combined way FEM to extract the effective dynamic and structural behaviour of the interiors, and the multibody method (MB) to determine the kinematic of secondary impacts and biomechanical parameters. It has to be pointed out that when one is not interested to internal stress and strain states coming from a dynamic phenomenon, a different method can be used, the multibody one, which is

very fast and efficient and which can be coupled with a FEM analysis when completing the study.

The steps followed to develop this activity are listed below:

1. obtaining by FEA the "pulse" necessary to initialize the multibody analysis: within this phase the dynamic behaviour of the interiors have been also evaluated;
2. obtaining by MB analysis the kinematic behaviour of passengers;
3. obtaining contact stiffness of the interiors by local FE analyses, which has been used to characterize the stiffness of the panels in the multibody environment;
4. simulating the secondary impacts in a multibody environment.

According to the EN15227 standard, the selected collision scenario has been the frontal impact between two similar vehicles (Fig. 21) at a speed of 36 km/h; such scenario has been modelled and analyzed, by using the explicit finite element code LS-Dyna®, as a collision of a single vehicle against a rigid barrier at a speed of 18 km/h.

Fig. 21. The FEM model of a train coach

The first phase of the analysis has regarded the estimation of the deformations of the vehicle as a whole, with the aim to evaluate the reduction of the occupants/driver survival space and the probable disengagement of the bogie wheels from the rail. Stated the respect of these standard requirements, the successive phase has regarded the analysis of the energies involved in the phenomenon (Fig. 22); the value of the initial kinetic energy of the vehicle is 851,250 J, which at the end of the impact is fully converted into internal energy of the system. It should be considered that the internal energy includes the elastic energy stored by the buffer spring, which is recovered in terms of kinetic energy during the "bounce" of the vehicle. As it can be seen in Fig. 22, about 50,000 J are absorbed in the first phase of the impact by the buffer; once the buffer spring has been fully compressed, about 600,000 J are absorbed by the two absorbers, proportionally to their characteristics.

The next analyzed resulting parameter is the acceleration, which in this case has been evaluated on the “rigid” pin linking the structure to the forward bogie. As it can be seen from the plot in the lower left of Fig. 22, which will be the "pulse" for the Multibody analysis, during the absorption of the impact energy by the buffer/absorbers group, the maximum acceleration value is about 5g, to grow up to about 15g when the frame is involved in the collision.

Finally, it has been evaluated the interface reaction between the vehicle and the barrier (Fig. 22): it is almost constant, with acceptable maximum value, until the frame is involved in the collision.

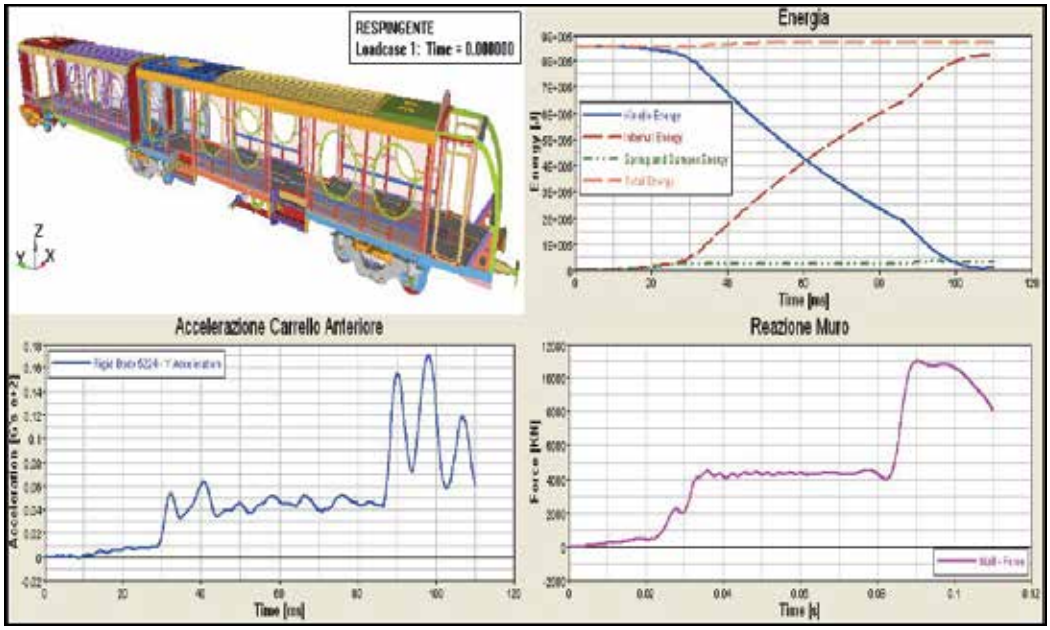


Fig. 22. Pulse evaluation from FEM analysis

The main objective of the work is to develop a complete multibody model of a critical area inside a train unit, including the model of an anthropomorphic dummy, which allows to develop fast simulations of secondary impact scenarios from which to obtain biomechanical results; moreover, by proceeding in this way, it is also possible to quickly evaluate the changes in biomechanical performances of the interiors that characterize different configurations (stiffness of the panels, thickness and arrangement of the reinforcement, etc.).
In order to characterize the contact reaction between the dummy and the interiors in a multibody environment, the panels are modelled as rigid bodies, but their impact surfaces react to the impact by following an assigned law of the reaction forces vs. displacement through the contact surface. This law must be evaluated either by considering experimental compression tests of the panel, or by developing a local finite element analysis by modelling the real properties of the materials of the panels.
The advantages in the use of this hybrid methodology are briefly described below:

- a full multibody model (free from FE surfaces) requires very short calculation time;
- the multibody model is a very flexible one, in which it is possible to change the "response of the material" by acting only on the characteristics of stiffness at the contact interface;
- the change in geometry of the multibody model is very simple and fast.

In Figs. 23 and 24, we show some images related to the preliminary multibody analysis performed by using Madymo® MB commercial code, by considering as perfectly rigid the surfaces representing all the components of the considered scenario. This analysis provides information about the kinematic of the secondary impacts involving a generic seated passenger (Dummy "Hybrid_III_95% ile") and a composite panel positioned in front of him. We also introduced the hypothesis that the effective stiffness of the impacted panels doesn't influence the relative kinematic between the panels and the passengers.

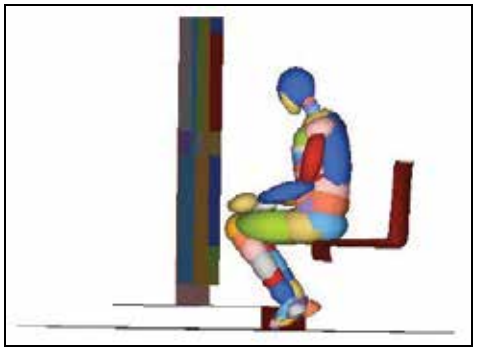


Fig. 23. 140 ms, the dummy breaks away from the seat; feet are blocked under the step and are not able to slide on the floor.

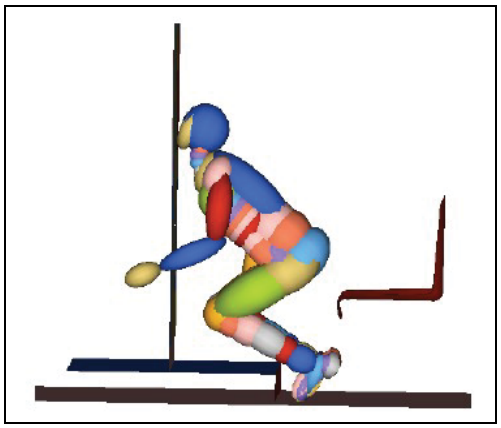


Fig. 24. 235 ms: the neck reaches the critical position: this is the maximum deflection

From this preliminary analysis it was possible to extract information about the exact areas of the panels interested from the impact with the passenger; the next step was to develop a explicit finite element analysis in order to evaluate the effective "contact stiffness" of these areas.

To evaluate by explicit FE analysis the effective stiffness of the interior panels it is not useful to consider a sub-model of the areas of the whole panel interested from the impact, because of the effective local stiffness depends on the effective boundary conditions, in terms of type and position of the constraint and of the stiffeners positioned beside the panel.

For what concern the dummy, in the finite element analysis it has been replaced by a series of rigid spherical bodies, with an opportune calibrated mass (19 kg for knee and 9.6 for the head) and with a specific speed (5 m/s), in order to obtain the same impact energy value. The impact areas were chosen considering the kinematic analysis made previously and in particular they have been chosen considering the knees and the head impact areas.

For every collision were considered 4 cases for the knees impacts, and 4 cases for the head impacts, two of which are showed in figures 25 and 26. From those analyses it has been possible to obtain the effective stiffness of the panels to set up the "contact stiffness" of the multibody ones.

Fig. 25. Comparison between contact forces in Dyna-Madymo in the head area

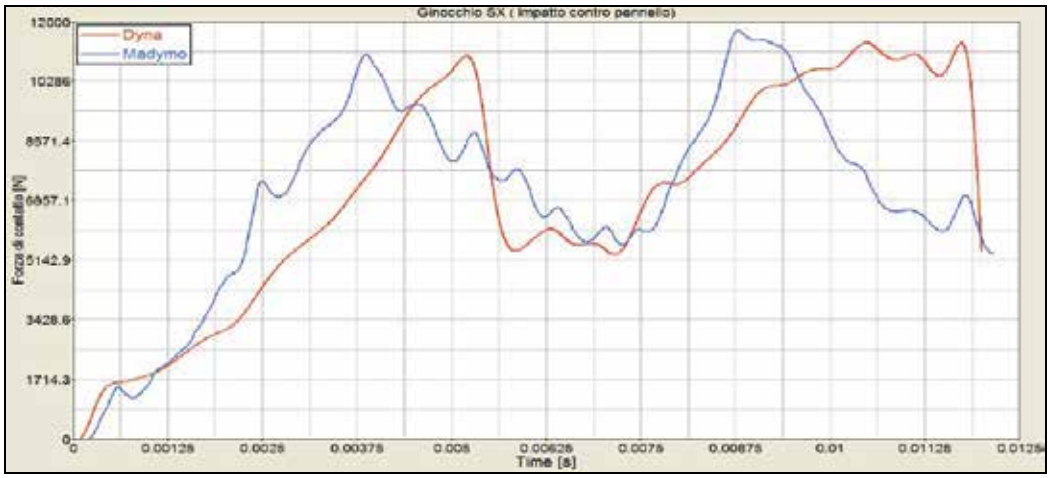


Fig. 26. Comparison between contact forces in Dyna-Madymo in the area of the SX knee

The thus obtained contact stiffness was used to characterize the panels and a multibody analysis was developed in order to obtain the biomechanical indexes. To evaluate different scenarios of secondary impact, some changes to the initial model were considered; the changes concern:

- replacing of the step on the floor by a ramp;
- changing the position of the seat by the maximum distance from the panel.

The results obtained from the full multibody analysis are reported in terms of the biomechanical indexes characterizing the impact scenarios described in the previous section. In Tab. I the biomechanical indexes related to the head are reported; in the last column the limit values are illustrated for each index.

Following the previous analysis it was possible study the 'best' configuration of the interiors in order to limit damages to passengers during the secondary impact; for example, it was possible to confirm that the best configuration was that where the step was no longer present and the seat had the maximum distance from the panels.

Biomechanical indexes	With step	Without step, Seat in a rearward position	Advanced position	Limit of value
Head Injury Criterion (36ms)	**5865.1**	**2866.0**	**2540.9**	500
Resultant head acceleration (3ms,cumulative)	**270.6** (m/s^2)	**1372.9** (m/s^2)	**1888.0** (m/s^2)	800 (m/s^2)
Normalized neck injury criteria	**2.402**	0.436	0.768	1
Viscous criteria	4.301E- 03(m/s)	4.100E-02 (m/s)	0.216 (m/s)	1(m/s^2)

Table I Numerical values of the obtained biomechanical indexes

5.3 Crashworthiness analyses in aeronautics

This work deals with a numerical investigation about the most reliable procedure to simulate, by finite element method, a sled-test to certificate aeronautical seat. These types of tests are mostly characterized by strongly dynamic effects, even if some evaluations about structural behaviour under quasi-static load conditions are required to certificate the seat. Generally, to develop numerical analyses of dynamic behaviour, explicit finite element algorithms are used; to develop quasi-static analyses both explicit and implicit methods could be suitable. Comparisons between results carried out by using both the methods have been developed, in terms of accuracy of results, calculation time and feasibility of preprocessing phase.

As a reference case we choose an archetype of a passenger seat of an helicopter which is comprised in the "Small Rotorcraft" category, as defined by EASA CS-27 standard. The numerical simulation refer to the "test 2 AS8049 SAE" which states that the seat (dummy included) is subjected at first to a displacement set such as to represent the effect of the deformation of the floor, in quasi-static conditions, then an assigned velocity is impressed to it and at last it is stopped according to a prescribed deceleration curve. It is then possible to identify two distinct phases in the test, the first being characterized by quasi-static phenomena (pre-crash) and the second one accompanied by largely dynamical phenomena (crash).

As the advantages of explicit FEM of the crash phase are well known, the attention was focused on the analysis methods of quasi-static phenomena which characterize pre-crash phase and which in our case are as follows:

- the introduction of rotation of the seat mounting to simulate the effect of the deformation of the aircraft floor;
- the positioning of the dummy and the simulation of the subsequent crushing of the set foam.

The aim of the whole procedure was to find out the most convenient analysis conditions to simulate pre-crash phase, for what refers to reliability of results, computational weight and user-friendliness of preprocessing.

According to AS 8049 SAE standard, a minimum of two dynamical tests is required to certificate the seat and the restraining system, which have both to protect the passenger in the crash phase. On the present work, the test no. 2 was simulated, which considers that a 12.8 m/s velocity is impressed to the seat, which is mounted on a sled, after subduing quasi-static deformations, and which is then stopped in 142 ms, according to a triangular deceleration profile. The inertia forces resultant is directed along a 10° direction with respect

to the longitudinal axis of the aircraft, because of the presence of the main component, which is directed along the longitudinal axis of the aircraft.
The effect of the floor deformation was simulated by applying assigned rotations to the links of the seat to the aircraft structure, which generally occurs through rails which are called "seat tracks". Pitch and roll beam angles, thus simulating the behaviour of the seat-tracks, are assumed to be 10°, and their direction is such as to simulate the hardest load condition.
The two rotations occur in 100 ms each, according two functions whose behaviour can be subdivided in three intervals, as follows:

- increasing velocity according a linear law, from 0 to 2.627 rad/sec;
- constant velocity, at 2.627 rad/sec;
- deceleration, according a linear law up to stop.

The procedure was carried by using the commercial code RADIOSS which let the user choose between explicit and implicit integration; that capability was very useful in this case, because explicit codes require very long runtimes when analyzing quasi-static conditions. In Fig. 27 the results are shown for both explicit and implicit analysis of the connection substructure between the seat and the floor, as appearing after a 10° rotation of the junction between the right leg and the floor; it can be seen that the results are almost the same for the two formulations.

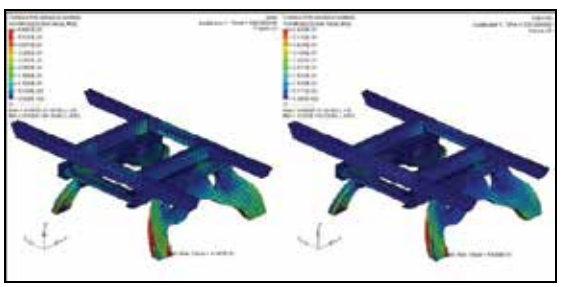


Fig. 27. Von Mises stress as obtained through the explicit (left) and implicit (right) methods

As the object of this paper was the evaluation of the behaviour of the seat, neglecting for the time being the analysis of the passenger, the latter could be simulated by means of a simplified dummy, which could be a rigid one, without joints, with the whole mass was concentrated in its gravity center. A second rigid body was introduced to simulate the whole structure of the seat, but for the elements which represent the two cushions; that behaviour doesn't invalidate the procedure and let reduce greatly the subsequent runtimes.
In the following Fig. 28 we represented the plot of the vertical displacement of the gravity center of the dummy and the kinetic energy of the system as functions of time; the max displacement (6.36mm) is the same for both formulations (implicit and explicit).
After the previous analyses, a complete run for the whole certification test was carried out through an explicit code. In Fig. 29 we have the plots of energy, velocity and acceleration which refer to the master nodes of the rigid elements which simulate the connection between the seat and the floor. For what refers to the kinetic energy, we can observe a point of discontinuity after 200 ms from the beginning of the test: it corresponds to the separation point between the quasi-static phase and that highly dynamic of crash phase. The peak value of kinetic energy appears just at the beginning of crash and amounts to 7680 J, i.e. the total energy of the whole system, whose mass is 93.75 kg, when its velocity is 12.8 m/s.

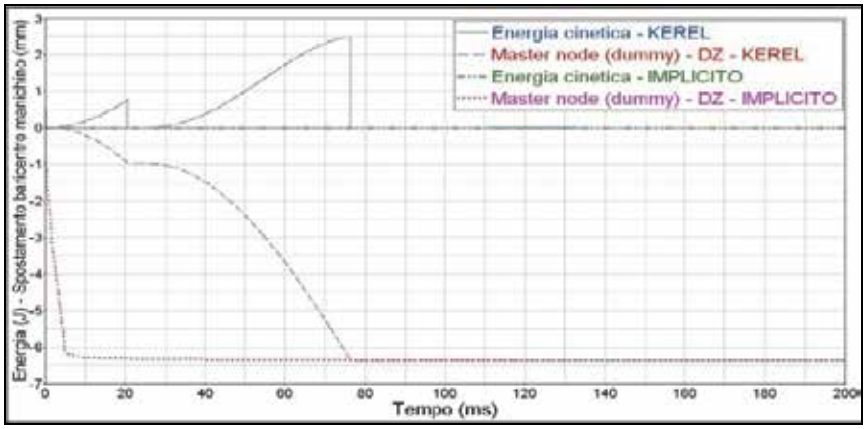


Fig. 28. Vertical displacement and Energy of the gravity center of the dummy

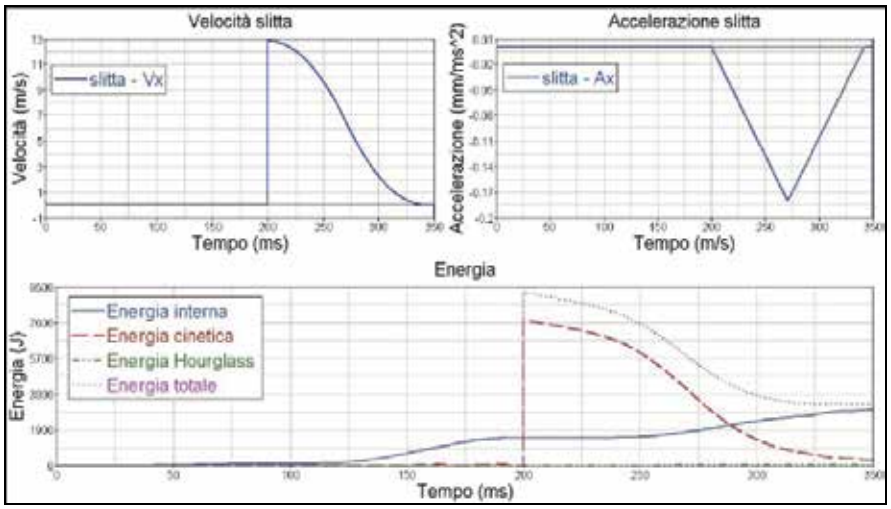


Fig. 29. Velocity and acceleration of the sled, with the absorbed Energy levels

6. Conclusions

Today available explicit codes allow the analyst to study very complex structures in presence of impulsive loads; the cases considered above show the degree of deepening and the accuracy which can be obtained, with a relevant gain in such cases as manufacturing, comfort and safety.

Those advantages are in any case reached through very difficult simulations, as they require an accurate modeling, very fine meshes and what is more relevant, a sound knowledge of the behaviour of the used materials in very particular conditions and in presence of high strain rates.

The continuous advances of computers and of methods of solution let us forecast in the near future a conspicuous progress, at most for what refers to the speed of processors and algorithms, what will make possible to perform more simulations, yet reducing the number of experimental tests, and to deal with the probabilistic aspects of such load cases.

7. References

Chang, J.M.; Tyan, T.; El-bkaily, Cheng, M.J.; Marpu, A.; Zeng, Q. & Santini, J. (2007), Implicit and explicit finite element methods for crash safety analysis, *Proceedings of World Congress Detroit, Michigan;* SAE, Warrendale, Pa.;

Clausen, A.H.; Hopperstad, O.S. & Langseth, M (2001). Sensitivity of model parameters in stretch bending of aluminium extrusions, *J. Mech. Sci.* Vol. 43, p. 427. doi:10.1016/S0020-7403(00)00012-6

Dias, J.P. & Pereira, M.S. (2004). Optimization methods for crashworthiness design using multibody models, *Composite & Structures*, vol 82, pp. 1371-1380, ISSN 0263-8223;

Drazetic, P.; Level, P.; Cornette, D.; Mongenie, P & Ravalard, Y. (1995). One-Dimensional modelling of contact impact problem in guided transport vehicle crash, *Int. J. Impact Engng*, vol 16/3, pp 467 – 478, ISSN 0734-743X;.

European Standard EN 15227 (2008). Railway applications - Crashworthiness requirements for railway vehicle bodies;

Fitzgerald, T.J. & Cohen, J.B. (1994). Residual Stresses in and around Rivets in Clad Aluminium Alloy Plates, *Materials Science & Technology*, vol. AI88, pp. 51-58, ISSN 0267-0836;

Kirkpatrick, S.W.; Schroeder, M. & Simons, J.W. (2001). Evaluation of Passenger Rail Vehicle Crashworthiness, *International Journal of Crashworthiness*, Vol. 6, No. 1, pp. 95-106, ISSN 1358-8265;

Kirkpatrick, S.W. & MacNeil,R.A. (2002). Development of a computer model for prediction of collision response of a railroad passenger car, *Proceedings of the 2002 ASME/IEEE Joint Rail Conference*, pp. 9 - 16, Washington, DC, April 23-25; ISBN 0-7918-3593-6.

Kradinov, V.; Barut, A.; Madenci, E. & Ambur, D.R. (2001). Bolted Double-Lap Composite Joints under Mechanical and Thermal Loading, *Int. J. of Solids & Structures*, vol. 38, pp. 7801-7837, ISSN 0020-7683;

Langrand, B.; Patronelli, L.; Deletombe, E.; Markiewicz, E. & Drazétic:, P. (2002). Full Scale Characterisation for Riveted Joint Design, *Aerospace Science & Technology*, vol. 6, pp. 333-342, ISSN 1270-9638;

Langrand, B.; Fabis, J.; Deudon, A. & Mortier, J.M. (2003). Experimental Characterization of Mechanical Behaviour and Failure Criterion of Assemblies under Mixed Mode Loading. Application to Rivets and Spotwelds, *Mécanique & Industries*, vol. 4, pp. 273-283, ISSN 1296-2139;

Madenci, E.; Shkarayev, S.; Sergeev, B.; Oplinger, O.W. & Shyprykevic, P. (1998). Analysis of composite laminates with multiple fasteners", *Int. J. of Solids & Structures*, vol. 35, pp. 1793-1811, ISSN 0020-7683;

Moon, Y.H.; Kang, S.S.; Cho, J.R. & Kim, T.G. (2003). Effect of tool temperature on the reduction of the springback of aluminum sheets, *J. Mater. Process. Technol.*, Vol 132 p. 365. doi:10.1016/S0924-0136(02)00925-1, ISSN 0924-0136;

Ryan, L.; Monaghan, J. (2000). Failure Mechanism of Riveted Joint in Fibre Metal Laminates, *J. of Materials Processing Technology*, vol. 103, pp. 36-43, ISSN 0924-0136;

Schiehlen, W. (2006). Computational dynamics: theory and applications of multibody systems, *European Journal of Mechanics A/Solids,* vol. 25, pp. 566–594, ISSN 0997-7538;

Schiehlen, W.; Guse, N. & Seifried, R. (2006). Multibody dynamics in computational mechanics and engineering applications, *Comput. Methods Appl. Mech. Engrg*, vol. 195, pp. 5509-5522, ISSN 0045-7825;

Severson, K.; Perlman, A.B. & Stringfellow, R. (2008). Quasi-static and dynamic sled testing of prototype commuter rail passenger seats, *Proceedings of the 2008 IEEE/ASME Joint Rail Conference, JRC2008*, ISBN 0791848124, April 22-23, 2008 Wilmington, Delaware, USA.

Urban, M.U. (2003). Analysis of the Fatigue Life of Riveted Sheet Metal Helicopter Airframe Joints, *Int. J. of Fatigue*, vol. 25, pp. 1013-1026, ISSN 0142-1123;

Zhang, J.; Park, J.H. & Atluri, S.N. (1997). Analytical Fatigue Life Estimation of Full-Scale Fuselage Panel, in: *Proc. of FAA-NASA Symposium on the Continute Airworthiness of Aircraft Structures, DOT/FAA/AR-97/2,* vol. 1, pp. 51-62, ISBN 87404-279-8, Atlanta, GA, August 28-29.

13

Inverse Methods on Small Punch Tests

Inés Peñuelas, Covadonga Betegón, Cristina Rodríguez and Javier Belzunce
University of Oviedo
Spain

1. Introduction

The characterization of the mechanical behaviour of structural materials, with the exception of material hardness, is a destructive procedure which requires direct extraction of test specimens from the component to analyse. Because this component needs to be operative, these specimens have to be as small as possible, in order not to affect the behaviour of the component and in order to allow easy reparation of the 'damaged' component. However, tests with miniaturized specimens are not defined in standards. Thus, the results obtained with these tests have to be interpreted in order to obtain the actual properties of the components from which the specimens have been extracted (Lucas et al., 2002). The small punch test (SPT) is very useful in all applications that require the characterization of the mechanical behaviour of structural materials or operational components without compromising their service (Lucon, 2001), as in the case of nuclear or thermal plants. Another application is the study of small testing zones. Thus, this test has been recently applied to the mechanical characterization of metallic coatings (Peñuelas et al, 2009) or the heat affected zone of welds (Rodriguez et al, 2009), which are practically impossible to characterize by means of the conventional mechanical tests.

Advance constitutive models frequently include parameters that have to be identified through numerical simulation of tests and mathematical optimization of variables, because they cannot often be directly measured in laboratory. In this paper, an inverse methodology for the identification of the mechanical and damage properties of structural steels has been developed. Thus, from the load-displacement curves obtained during the non-standard SPT, the mechanical and damage properties will be obtained. Moreover, this methodology also allows simulating the SP test with numerical methods.

Structural steels may exhibit creep behaviour and behave according to the Hollomon's law ($\sigma = K\,\varepsilon_p^n$). Besides, ductile fracture of metallic materials involves micro-void nucleation and growth, and final coalescence of neighbouring voids to create new surfaces of a macro-crack. The ductile failure process for porous materials is often modelled by means of the Gurson model (Gurson, 1977), which is one of the most widely known micro-mechanical models for ductile fracture, and describes the progressive degradation of material stress capacity. In this model, which is a modification of the von Mises one, an elastic–plastic matrix material is considered and a new internal variable, the void volume fraction, f, is introduced. Although the original Gurson model was later modified by many authors, particularly by Tvergaard and Needleman (Tvergaard, 1981; Tvergaard, 1982; Tvergaard & Needleman, 1984), the resultant model is not intrinsically able to predict coalescence, and is only capable of

simulating micro-void nucleation and growth. This deficiency is solved by introducing an empirical void coalescence criterion: coalescence occurs when a critical void volume fraction, f_c, is reached (Tvergaard, 1982; Koplik & Needleman, 1998; Sun et al. 1992). Combining these models, it is possible to simulate the behaviour of materials from the elastic behaviour until their total fracture. The macromechanical and micromechanical parameters relate with different zones of the load-displacement curve obtained with the SPTs. These zones will be described below.
In the inverse procedure considered here, most data are pseudo-experimental data, that is, they are obtained from the numerical simulation of the test for a prescribed set of material parameters. Notwithstanding, many real experimental data are also considered in order to validate the numerical model and the inverse methodology developed.

2. Inverse methodology

The methodology used in this paper is based on inverse methods (Stravroulakis et al., 2003), design of experiments (Kuehl, 2000; Montgomery, 1997), numerical simulations of tests, least-squares polynomial regression for curve fitting and evolutionary genetic algorithms (Deb, 2001; Seshadri, 2006). Inverse problems lead to difficult optimization problems whose solutions are not always straightforward with current numerical optimization techniques. Therefore, one should consider semi-empirical methods and experimental testing techniques as well (Bolzon et al., 1997). Design of experiments (DOE) is the methodology of how to conduct and plan experiments in order to extract the maximum amount of information in the fewest number of runs. The statistical experiment designs most widely used in optimization experiments are termed response surface designs (Myers & Montgmomery, 1995). In addition to trials at the extreme level settings of the variables, response surface designs contain trials in which one or more of the variables is set at the midpoint of the study range (other levels in the interior of the range may also be represented). Thus, these designs provide information on direct effects, pair wise interaction effects and curvilinear variable effects. Properly designed and executed experiments will generate more precise data while using substantially fewer experimental runs than alternative approaches. They will lead to results that can be interpreted using relatively simple statistical techniques. If there are curvilinear effects the factorial design can be expanded to allow estimation of the response surface. One way to do this is to add experimental points. The central composed design uses the factorial design as the base and adds what are known as star points. Special methods are available to calculate these star points, which provide desirable statistical properties to the study results.
In the inverse methodology, for the numerical and experimental tests, the different zones of the load-displacement curve have to be fitted. Data fitting is usually done by means of an error minimization technique, where the distance between parameterized predictions of the mechanical model (parameterized by the unknown parameters) and measurements of the corresponding experiment is minimized. This formulation is known as an output error minimization procedure for the inverse problem (Stravroulakis et al., 2003). In order to choose the best fitting model for all of them, for each fitting model, different statistical coefficients have been analysed:

1. The coefficient of multiple determination, also called proportion of variance explained R^2, that indicates how much better the function predicts the dependent variable than

just using the mean value of the dependent variable (the closer to 1.0 (100%), the best the function predicts the observed data);
2. The adjusted coefficient of multiple determination R_a^2 that is an R^2 statistic adjusted for the number of parameters in the equation and the number of data observed (the closer to 1.0 the best the function predicts the observed data);
3. The Durbin–Watson statistic, used to detect the presence of autocorrelation in the residuals from the regression analyses (a value less than 0.8 usually indicates that autocorrelation is likely (autocorrelation should be avoid));
4. The t-ratio, that is a measure of the likelihood that the actual value of the parameter is not zero (the larger the absolute value of t, the less likely that the actual value of the parameter could be zero) and
5. The prob(t) value that is the probability of obtaining the estimated value of the parameter if the actual parameter value is zero (the smaller the value of prob(t), the more significant the parameter and the less likely that the actual parameter value is zero).

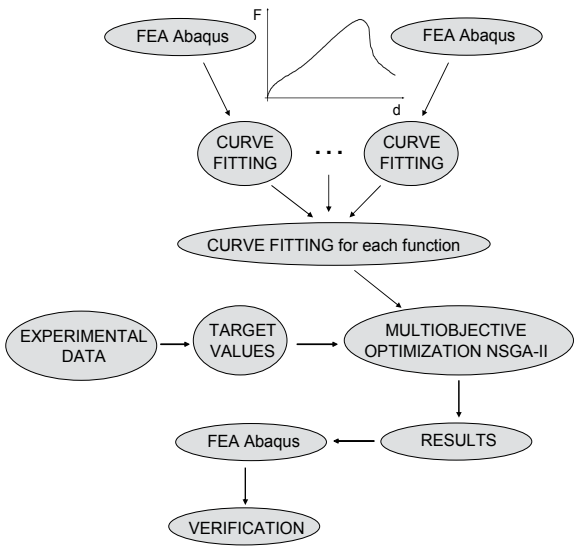


Fig. 1. Scheme for the inverse procedure

Inverse procedure finishes with the determination of the set of variable values that are associated to certain target values, obtained from the load-displacement curve of a laboratory SPT. That is, it have to be searched the set of variable values that simultaneously minimize a certain number of objective functions. This is a multiobjective optimization problem that can be solved using different procedures. In this paper, the Pareto front has been obtained by means of the evolutionary genetic algorithm NSGA-II (Seshadri, 2006). Pareto front produces non-dominated set of solutions with regard to all objectives and all solutions on the Pareto front are optimal. Besides, NSGA-II is non-domination based genetic algorithm which incorporates elitism (only the best individuals are selected) and that does not requires choosing *a priori* sharing parameters. This algorithm is run in MATLAB. First of all the population is initialized based on the problem range and constraints if any. This population is sorted based on no domination (an individual is said to dominate another if

the objective functions of it, is no worse than the other and at least in one of its objective functions it is better than the other). Once the non-dominated sort is complete, a crowding distance, that is a measure of how close and individual is to its neighbours, is assigned. Parents are selected from the population by using binary tournament selection based on the rank and crowding distance. The selected population generates offspring from crossover and mutation operators. The population with the current population and current offspring is sorted again based on non-domination and only the best N individuals are selected, where N is the population size. Fig. 1 shows the scheme for the inverse procedure used for the material characterisation.

3. Small punch test (SPT)

By virtue of the small size of the specimens required for testing, the Small Punch Test can be considered a non-destructive test. Usually, the specimens used for the SPT are square plates of 10 × 10 mm² and just 0.5 mm thickness, although lower or higher thickness can also be used. In comparison with other non-destructive techniques such as ultrasonic or magnetic techniques and X-Rays, that are based on indirect measures for the determination of the above mentioned properties, the SPT allows obtaining directly the main mechanical properties of the materials.

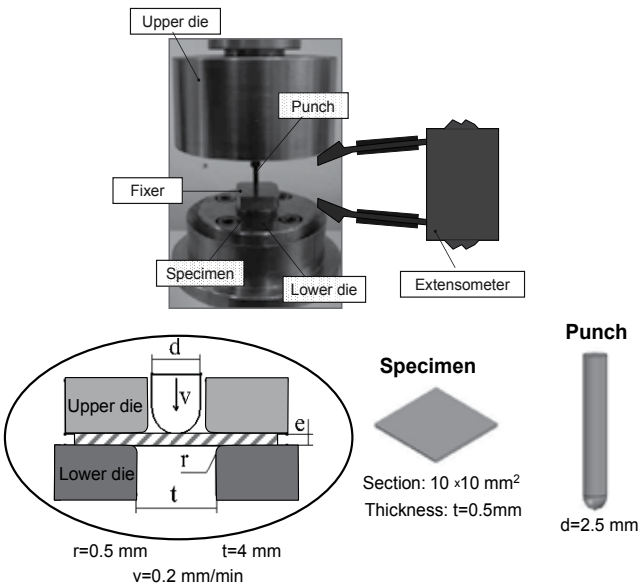


Fig. 2. Dispositive and geometry of the small punch test

In laboratory, the SPTs have been carried out with a low speed tensile test machine. Test consists of fixing the periphery of the specimen, embedding it between two dies (upper and lower dies) by means of four screws and a tightening torque of 2 N·m, and then deforming the specimen until its fracture by means of a small semi-spherical punch with a head of 2.5 mm of diameter. The test is speed controlled with a punching speed v = 0.2 mm/min. In this way, the specimen is bounded to deform quasi-statically inside a 4 mm diameter hole (biaxial expansion) up to failure (Fig. 2). The data sampling rate during the experiment is 20

samples/s. Moreover, the test is finalized when load decreases the 50% of the maximum load.

By means of an extensometer, the displacement of the punch is obtained, and after correction of the flexibility of the testing device, the displacement of the central point of the specimen is calculated. Thus, from test is obtained the characteristic curve of material. This curve represents the force exerted from punch against the specimen (i.e. the load reaction) versus the displacement of the punch (Fig. 3). In the case of ductile materials, six different zones can be distinguished in these load-displacement curves obtained by means of the SPTs: zone I (elastic deformation), zone II (elastoplastic transition), zone III (generalized plastic deformation), zone IV (plastic instability and fracture initiation), zone V (fracture softening zone) and zone VI (final fracture).

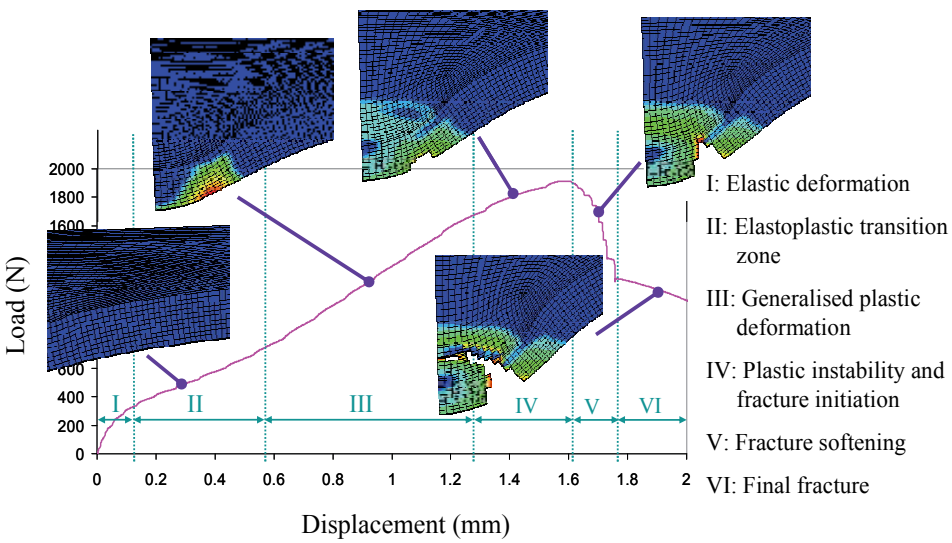


Fig. 3. Load-displacement curve for the SPT and Finite Element simulation at each zone

4. Numerical simulation of the SPT

Different models have been developed in order to reproduce the SPTs by means of numerical methods. These models were compared with the aim of choosing the optimum model from the point of view of the relation between the precision and the computational cost. The numerical simulations have been carried out with the finite element commercial code ABAQUS (ABAQUS 6.7, 2008). In order to simulate the fracture behaviour of isotropic and anisotropic materials, two different meshes have been used (2D and 3D meshes, respectively). As it was pointed out before, the specimens for laboratory are squared specimens. However, because the hollow between the die and the specimen is a cylinder, the problem can be considered axisymmetric in the isotropic model, and the model can be solved by 2D axisymmetric meshes. Besides, for isotropic materials the 3D model has been compared with the axisymmetric one (2D) in order to justify the use of the axisymmetric model for the sake of simplicity. In the 2D-Axysim model, the specimens were discretised by means of an axisymmetric mesh of four-node reduced integration hybrid elements.

Notwithstanding, since many structural steels are obtained from lamination processes, they exhibit anisotropic behaviour. In these cases, three-dimensional meshes which reproduced a quarter of the specimen were used (Fig. 4). Although geometries of Fig. 4 appear to be different, the applied boundary conditions allow using both of them for isotropic materials. In this figure, upper die is not represented in order to improve the visualization of the model. In all cases, die and punch were modelled as rigid bodies. Besides, contact between surfaces, quasistatic analysis and large displacements were taken into account.

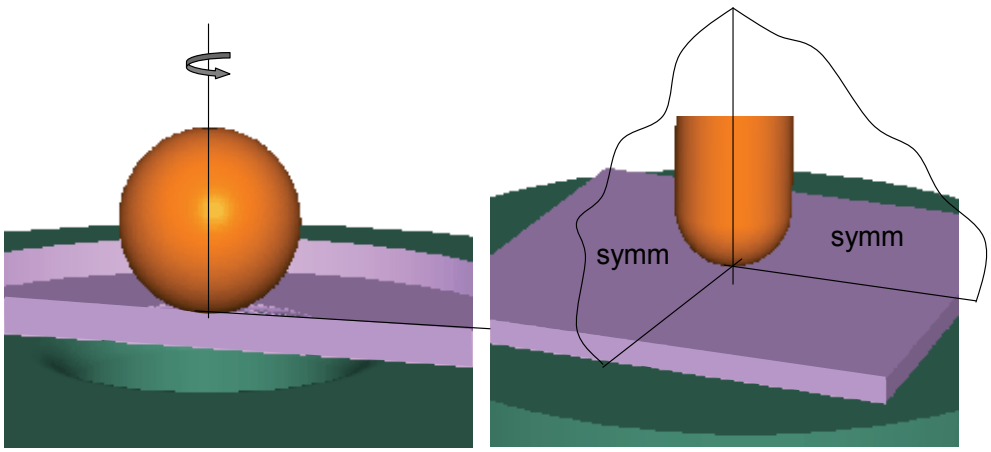


Fig. 4. Axysimmetric and Three-dimensional models used for the simulation of the SPT

From sensitivity analyses, it is observed that the elastic and elastoplastic transition zones (zones I and II of the load-displacement curve) are enough to characterize the macromechanical behaviour of steels that exhibit creep behaviour and follow the Hollomon's law ($\sigma=K\ \varepsilon_n^p$), whereas the remaining zones allow to characterize the micromechanical behaviour of the material and the coefficient of friction to be used in simulations.

In order to choose a value for the coefficient of friction, different simulations of known materials have been carried out. A good approximation has been obtained with $\mu = 0.1$, which is also an adequate value for steel–steel contact under partial lubrication. In the case of tests carried out with no lubrication, better results have been obtained with $\mu = 0.25–0.35$. These values have been obtained by comparing the experimental curve for an already known material (characterized by means of standard tests) with numerical ones obtained by means of the test simulation of this material with different values of coefficient of friction.

To describe the evolution of void growth and subsequent macroscopic material softening, the yield function of Gurson modified by and Tvergaard and Needleman (Tvergaard & Needleman, 1984) was used in this work. This modified yield function is defined by an expression in the form

$$\Phi\left(q,p,\bar{\sigma},f\right)=\left(\frac{q}{\bar{\sigma}}\right)^2+2\cdot q_1\cdot f^*\cdot\cosh\left(-\frac{3\cdot q_2\cdot p}{2\cdot\bar{\sigma}}\right)-\left(1+q_3\cdot f^{*2}\right)=0 \qquad (1)$$

where $\bar{\sigma}$ is flow stress of the matrix material which relates with the equivalent plastic strain, f is the current void volume fraction, $p=-\sigma_m$ with σ_m the macroscopic mean stress and q is the macroscopic von Mises effective stress given by

$$q = \sqrt{\frac{3}{2} \cdot \left(S_{ij} \cdot S_{ij} \right)} \tag{2}$$

where S_{ij} denotes the deviatoric components of the Cauchy stress tensor. Constants q_1, q_2 and q_3 are fitting parameters introduced by Tvegaard (Tvergaard, 1981; Tvergaard, 1982) to provide better agreement with results of detailed unit cell calculations. The modified void volume fraction, f^*, was introduced by Tvergaard and Needleman (Tvergaard & Needleman, 1984) to predict the rapid loss in strength that accompanies void coalescence, and is given by

$$f^* = \begin{cases} f & \text{si} \quad f \le f_c \\ f_c + \dfrac{f^*_u - f_c}{f_F - f_c} \cdot \left(f - f_c \right) & \text{si} \quad f > f_c \end{cases}, \tag{3}$$

where f_c is the critical void volume fraction, f_F is the void volume fraction at final failure which is usually $f_F=0.15$ and $f^*_u=1/q_1$ is the ultimate void volume fraction.

The internal variables of the constitutive model are $\bar{\sigma}$ and f. Thus the evolution law for the void volume fraction is given in the model by an expression in the form

$$\dot{f} = \dot{f}_{growth} + \dot{f}_{nucleation} \tag{4}$$

The void nucleation law implemented in the current model takes into account nucleation of both small and large inclusions. The nucleation of larger inclusions is stress controlled, and it is assume that larger inclusions are nucleated at the beginning of the plastic deformation, being considered as initial void volume fraction. The nucleation of smaller inclusions is strain controlled and, accordingly to Chu and Needleman (Chu & Needleman, 1980) the nucleation rate is assume to follow a Gaussian distribution, that is

$$\dot{f}_{\text{nucleation}\ \text{small particles}} = A \cdot \dot{\bar{\varepsilon}}^p \tag{5}$$

where $\dot{\bar{\varepsilon}}^p$ is the equivalent plastic strain rate, and

$$A = \frac{f_n}{S_n \cdot \sqrt{2 \cdot \pi}} \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{\bar{\varepsilon}^p - \varepsilon_n}{S_n} \right)^2 \right) \tag{6}$$

where S_n is the standard deviation, ε_n is the mean strain and f_n is the void volume fraction of nucleating particles.

The growth rate of the existing voids can expressed as a function of the plastic strain rate in the form

$$\dot{f}_{growth} = \left(1 - f \right) \cdot \dot{\varepsilon}^p_{kk}, \tag{7}$$

where $\dot{\varepsilon}^p_{ij}$ is the plastic strain rate tensor.

5. Model calibration and sensitivity analysis

Prior to the inverse procedure is the direct adjustment of the numerical simulation and the experimental test for a small number of materials previously characterized by standard tests. That is the model calibration and it requires the determination of the unknown parameters of the model, especially of the ones relevant to defects and damage, by comparing the results of the model with experimental measurements. Afterwards, the load-displacement curves obtained from laboratory SPT and from FE simulation of the test for a material previously characterized from standard specimens, are compared. Fig 5 shows the qualitative comparison of the experimental and numerical deformation shape and fracture zones of the axisymmetric model at the final fracture of the specimen.

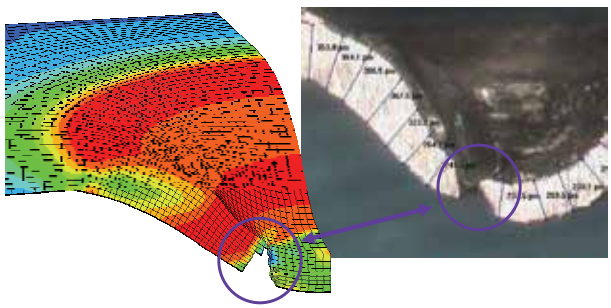


Fig. 5. Comparison of the experimental and numerical deformation shape and fracture zones of a SPT specimen

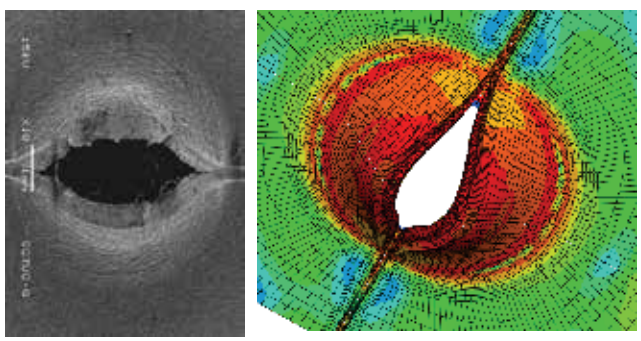


Fig. 6. Comparison of the experimental and numerical deformation shape and fracture zones of a notched SPT specimen

Moreover, Fig. 6 shows the comparison of deformation and overall appearance of the fracture zone obtained by a laboratory test and the numerical simulation, for SPT specimen with a longitudinal notch. In the case of notched specimens, 3D models has been used. It has been found very good correlation between tests and simulations, not only for the un-notched specimens but also for the notched specimens.

After setting the model, and before beginning the process of characterization, it is necessary to study which variables influence each of the zones of the load-displacement curve. For this purpose, several numerical simulations have been carried out. Fig. 7 shows the material parameters (variables to determine) that affect each zone of the load-displacement curve, obtained by means of SP tests.

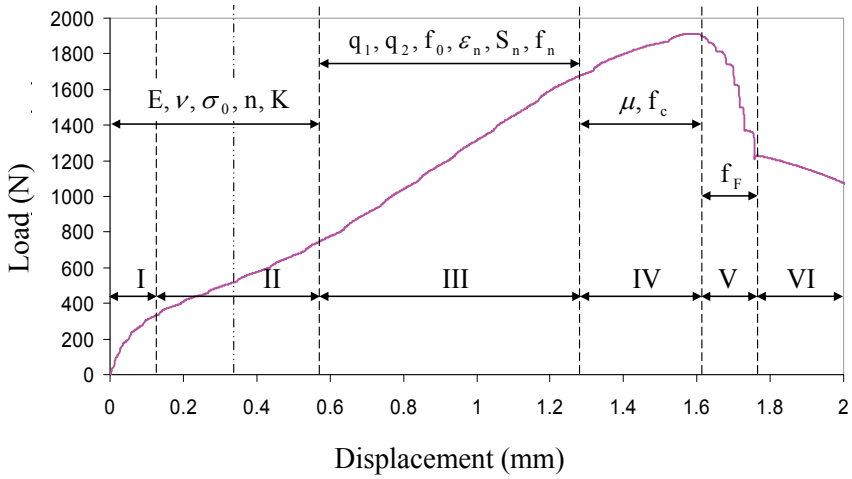


Fig. 7. Load-displacement curve of the SPT and parameters that affect each zone

From sensitivity analyses (±10%), it has been shown that load-displacement curves are very sensitive to variations in n and K (along the entire curve) and less sensible to variations in σ_0 (which mainly affects zone II). Moreover, since the SPT specimens reach the elastoplastic regime in the early stages of testing, the effect of Young's modulus is very small, so that E can be considered a constant reference value for all materials tested (analysed). Although the thickness of the specimen is a variable that has considerable influence on the load-displacement curve, in order to characterize the material is desirable using constant thickness. Therefore in Figure 7 is not shown the variable thickness-of-the-specimen. On the other hand, since the database has been obtained from pseudo-experimental data (numerical simulations), the technical problem of cutting all the specimens to the same small thickness (0.5 mm) is eliminated. Thus, for all simulations has been considered a fixed thickness.

6. Characterization methodology and results

As it was pointed out before, prior to the inverse procedure is the model calibration and the sensitivity analysis for the main variables. Afterwards, the inverse characterization scheme is applied. The complete material characterization requires the determination of a high number of parameters: coefficient of friction (μ), Young's modulus or elastic modulus (E), Poisson's ratio (ν), yield stress (σ_0), strain hardening exponent (n), Hollomon's factor (K), fitting parameters introduced by Tvergaard and Needleman for the GTN yield potential (q_1, q_2 and q_3), initial void volume fraction (f_0), mean strain in the Gaussian distribution of the nucleation rate (ε_n), standard deviation in the Gaussian distribution of the nucleation rate (S_n), void volume fraction of nucleating particles in the Gaussian distribution of the nucleation rate (f_n), critical void volume fraction (f_c) and void volume fraction at final failure (f_F). However, some of them can be obtained from literature or from previous works. This is the case for the μ, E, ν, q_1, q_2, q_3, f_0, S_n parameters. For metallic materials (structural steels) usual values of these constants are: E=2e5 MPa, ν = 0.3, q_1 = 1.5, q_2 = 1.0, $q_3 = q_1^2$ = 2.25 and S_n = 0.01 (small values of S_n relate to quite homogeneous materials). From metallographic observation of experimental specimens, the initial porosity has been considered f_0 = 0. And finally, from previous adjustments μ=0.1. Once the previous parameters are set, the number

of parameters to determine has been strongly reduced from 15 to 7: σ_0, n, K, ε_n, f_n, f_c and f_F. The first three parameters are macromechanic ones, the rest are micromechanic parameters for the damage model.

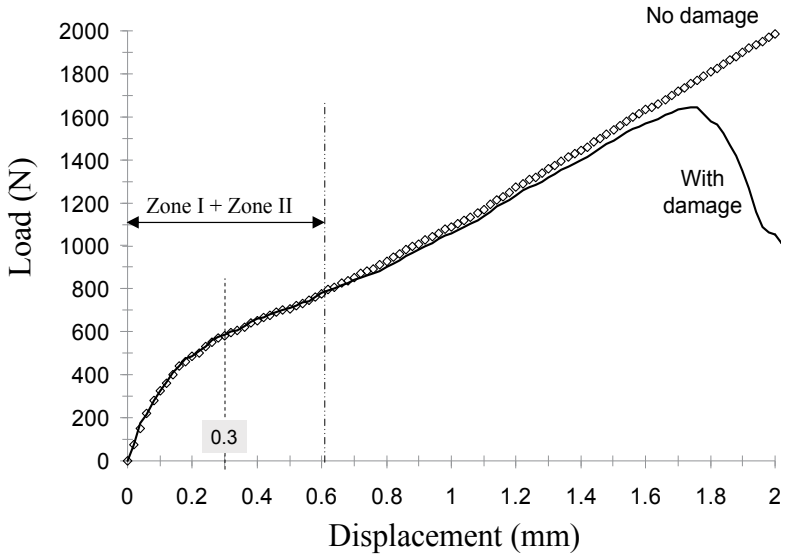


Fig. 8. Comparison of the Load-displacement curves with and without taking into account the damage of material

Since the damage parameters have no influence in the elastic and elastoplastic transition zones of the load-displacement curve (zones I and II), is possible to uncouple the macromechanical and the micromechanical characterizations. For this reason two different phases have been used for the macro- and micromechanical characterizations. First of all, the material has been macromechanically characterized by means of the analysis of zones I and II of the load-displacement curve. Then, the micromechanical parameters for the previously macro-characterized material have been determined using the remaining zones of the curve. Figure 8 shows the comparison between two numerical simulations for the same material with and without consideration of the damage model.

6.1 Macromechanical characterization

All inverse procedure requires a sufficiently large number of experimental data or pseudo-experimental data (numerical simulations). These data consist on sets of input variables for the macromechanical characterization (E, ν, σ_0, n, K) and output data obtained from the load-displacement curves. As it was pointed out before, the elastic modulus and the Poisson's ratio can be considered forehand known. Thus for a certain fixed values of the elastic modulus E = 2e5 MPa and the Poisson's ratio ν = 0.3, different combinations of (σ_0, n, K) have to be defined. In this paper, two different types of input variables have been taken into account. On the one hand, the design of experiments has been applied in order to define a small set of tests to simulate (15 tests). Thus, in order to identify the values of these sets of variables to simulate, it has been used design of experiments central composed centred on body, based on quadratic response surfaces. On the other hand, a wide battery of numerical

simulations (180 simulations) has been used not only to characterize the material macromechanically but also to quantify the effect of the simplifications inherent to the design of experiments. Besides, this battery is suitable for a wide range of structural steels. In both cases, the input variables vary within the following ranges

$$\sigma_0 = 200 - 700 \text{ MPa}, \; n = 0.1 - 0.3 \tag{8}$$

$$K = \begin{cases} 1.5 \cdot \sigma_0 - 3.5 \cdot \sigma_0 & \text{if } 0.1 \le n < 0.15 \\ 2.0 \cdot \sigma_0 - 4.0 \cdot \sigma_0 & \text{if } 0.15 \le n < 0.2 \\ 2.5 \cdot \sigma_0 - 4.5 \cdot \sigma_0 & \text{if } 0.2 \le n < 0.25 \\ 3.0 \cdot \sigma_0 - 5.0 \cdot \sigma_0 & \text{if } 0.25 \le n < 0.3 \end{cases} \tag{9}$$

In the case of using the battery of numerical simulations, the maximum variation of (σ_0, n) is $\Delta(\sigma_0, n)_{max}$=(50 MPa, 0.01).

In the design of experiments, it was considered a new variable K^* in order to correctly define the sets of values for simulation. This variable K^* varies from 1.5 to 3.5 and is given by

$$K^* = \left(\frac{K}{\sigma_0} - 0.5 \cdot i \right) \tag{10}$$

where i is defined by

$$i = \begin{cases} 0 & \text{if } 0.1 \le n < 0.15 \\ 1 & \text{if } 0.15 \le n < 0.2 \\ 2 & \text{if } 0.2 \le n < 0.25 \\ 3 & \text{if } 0.25 \le n < 0.3 \end{cases} \tag{11}$$

The output data were obtained from the curve fitting of zones I and II of the load-displacement curve in a two stage procedure which consists of:

1. First, fixing the range of displacement for the analyze. For all the structural steels simulated (180 steels with mechanical properties varying within the ranges defined before), a displacement value that has been proved to provided good results is 0.3 mm.
2. Then, adjusting the zone I and part of the zone II with an unique mathematical law. A commercial software, DataFit (DataFit 8.2, 2009) has been used for this purpose. The best fitting model is chosen by analyzing the different statistical coefficients of the different models. From the analysis of the different statistical coefficients of the different models, the best fitting model has been chosen. This consists in a exponential law in the form y=exp(a+b/x+c·Ln(x)), where y corresponds to load and x correspond to displacement. Fig. 9 shows this curve fitting for a generic material. In this way, the three output data obtained from each set of input data are the factors a, b, c, which depend on the three variables to determine, that is a=a(σ_0, n, K), b=b(σ_0, n, K) and c=c(σ_0, n, K).

Each of these functions is postulated as a polynomial model (Cuesta et al., 2007), being necessary determining its order. The higher this order, the bigger the number of coefficients to determine. Thus, in a second-order model the number of coefficients to determine is 10; in a third-order model is 20 and in a fourth-order model is 31. By the comparison of the numerical results obtained by the method of least-squares, and polynomial regressions of

orders two, three and four, it has been chosen to use the following models for the functions a,b,c: second-order models in case of using DOE for simulations and third-order models in case of using the battery of simulations, since they allow to reach good-enough adjustments using a relatively small number of coefficients. Table 1 gives detail of the R_a^2 values for each function a, b, c obtained with models of different orders. From this table can be observed that the adjusted coefficient of multiple determination is much higher for the battery of numerical simulations (180 simulations) than for the design of experiments (15 simulations). Besides, all the regressions used are very significant and the proportion of variance of a, b, c, explained are 99.7%, 95.6% and 98.4%, respectively.

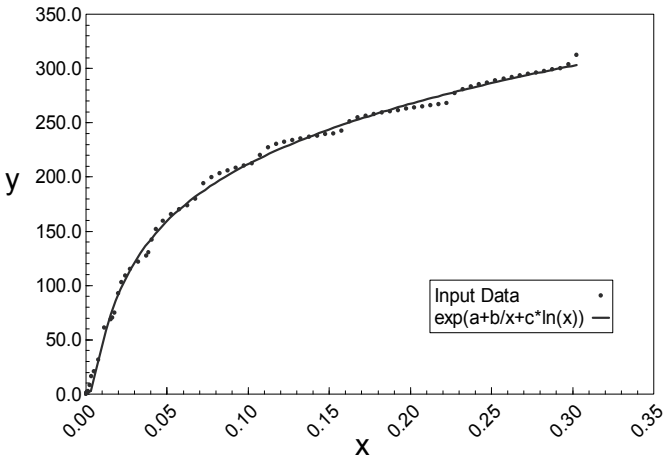


Fig. 9. Exponential adjustment of the Load-displacement curve in zone (I+II) until d=0.3 mm

Moreover, it has been carried out sensitivity analyses within a ±10% variation of the factors a, b, c of the exponential law, in order to analyse their effect on the load-displacement curve. These analyses show that the influence of the function a on the exponential law is enormous, the influence of c is notable and the influence of b is not important.

	Design of experiments	Battery of numerical simulations		
	2^{nd^o} order	2^{nd} order	3^{er} order	4^{th} order
a	0.981	0.986	0.997	0.999
b	0.889	0.934	0.951	0.960
c	0.907	0.961	0.982	0.987

Table 1. R_a^2. coefficients for functions a, b and c

In case of using design of experiments, the second order polynomial models for functions a, b and c can be write by expressions in the form

$$g(\sigma_0, n, K) = g_0 + g_1 \cdot \sigma_0 + g_2 \cdot n + g_3 \cdot K + g_{11} \cdot \sigma_0^2 + g_{22} \cdot n^2 + g_{33} \cdot K^2 + g_{12} \cdot \sigma_0 \cdot n + g_{13} \cdot \sigma_0 \cdot K + g_{23} \cdot n \cdot K \quad (12)$$

where $g(\sigma_0, n, K)$ correspond to $a=a(\sigma_0, n, K)$, $b=b(\sigma_0, n, K)$ and $c=c(\sigma_0, n, K)$.
Similarly, in the case of using the battery of numerical simulations, the third-order polynomial models for each function can be expressed in the form

$$g(\sigma_0, n, K) = g_0 + g_1 \cdot \sigma_0 + g_2 \cdot n + g_3 \cdot K + g_{11} \cdot \sigma_0^2 + g_{22} \cdot n^2 + g_{33} \cdot K^2 + g_{12} \cdot \sigma_0 \cdot n + g_{13} \cdot \sigma_0 \cdot K + g_{23} \cdot n \cdot K + g_{123} \cdot \sigma_0 \cdot n \cdot K + g_{112} \cdot \sigma_0^2 \cdot n + g_{113} \cdot \sigma_0^2 \cdot K + g_{122} \cdot \sigma_0 \cdot n^2 + g_{223} \cdot n^2 \cdot K + g_{133} \cdot \sigma_0 \cdot K^2 + g_{233} \cdot n \cdot K^2 + g_{111} \cdot \sigma_0^3 + g_{222} \cdot n^3 + g_{333} \cdot K^3 \quad (13)$$

Coefficients bijk have been obtained using the commercial software DataFit with regularized input values (σ_0, n, K) varying within the range [0, 1]. Values obtained for a 99% confidence interval are shown in Table 2.

	a	b	c
g_0	5.71048018	-0.00829258	0.25978748
g_1	0.00191446	-0.01508759	-0.36809281
g_2	-0.60585504	-0.00051838	-0.0886477
g_3	6.75259819	0.02658378	1.38774367
g_{11}	1.45337999	-0.00480519	0.33279479
g_{22}	-0.17753184	-0.0056283	-0.11501128
g_{33}	-2.88944339	-0.09971593	-1.81039397
g_{12}	0.00491019	-0.02896433	-0.48577964
g_{13}	-5.52592006	0.06567327	0.1338645
g_{23}	-1.12143967	0.04558742	0.55393563
g_{123}	-5.13987844	-0.06643944	-2.1955321
g_{112}	1.11391405	0.02857506	0.75839976
g_{113}	-1.44430745	-0.10182755	-1.94692201
g_{122}	1.40276526	0.00766641	0.38699445
g_{223}	-2.58686787	-0.01189609	-0.60625958
g_{133}	9.53781127	0.14445698	4.01278162
g_{233}	6.71653262	0.04137906	1.80579095
g_{111}	-0.47228307	0.01499177	0.10613477
g_{222}	0.42619501	0.00050424	0.0751837
g_{333}	-5.67785877	-0.04182362	-1.71486231

Table 2. g_{ijk}. coefficients for the third- order models for functions a, b and c

Finally, the inverse procedure finishes with the multiobjective optimization. That is, with the determination of the set of values (σ_0, n, K) that are associated to target values, which were obtained from the load-displacement curve of a specific laboratory small punch test. In our case, a_{target}=-6.097034, b_{target}=0.009365 and c_{target}=0.283507. Therefore, it have to be searched the set of variable values that simultaneously minimize three target (objective) functions: $(a - a_{target})$, $(b - b_{target})$ and $(c - c_{target})$. This multiobjective optimization problem has been solved using the evolutionary genetic algorithm NSGA-II, which has been run in MATLAB (MATLAB, 2006). The input arguments for the function nsga_2, are the population size and

number of generations. In this paper, the population size has been set to 200 and the number of generations has been set to 100. Since the algorithm incorporates elitism, only the best N individuals are selected, where N is the population size. The process repeats to generate the subsequent generations (100 generations). With this procedure the Pareto front is obtained, and it is represented in the space of functions [(a − a_{target}), (b − b_{target}),(c − c_{target})]. Fig. 10 shows the Pareto front in the space of functions for the target values.

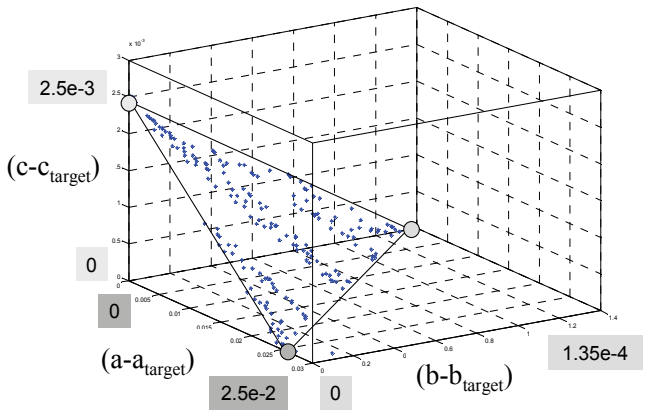


Fig. 10. Pareto front (Zone I+II) in the space of functions for the target values

As it was pointed out before, Pareto front produces non-dominated set of solutions with regard to all objectives and all solutions on the Pareto front are optimal. Furthermore, sensitivity analyses in functions a, b and c has shown that the variable that affects more the load-displacement curve (that is, the result) is variable a. As a result, from all the possible solutions that form the Pareto front, should be chosen those that show lower values of function objective (a − a_{target}). Fig. 11 shows the Pareto front in the space of solutions for the target values. Within this values it has been chosen one in the zone with higher population density of the solution space (σ_0, n, K), and it has been called the calculated set of variables (σ_0, n, K)$_{calculated}$. In order to verify its 'goodness', it has been compared with the values of the variables (σ_0, n, K) obtained by means of standard laboratory tests (traction test), which have been called the known values (σ_0, n, K)$_{known}$.

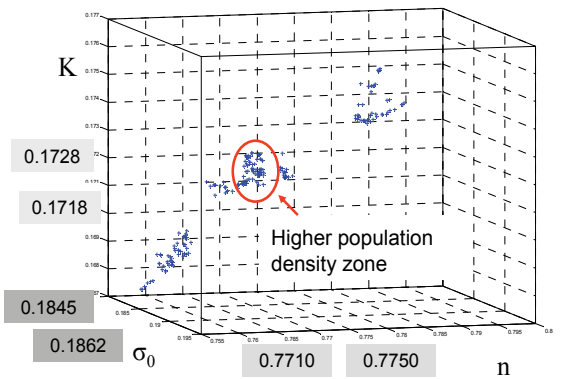


Fig. 11. Pareto front in the space of solutions

Moreover, the numerical simulation of the resulting values $(\sigma_0, n, K)_{calculated}$ has been carried out in order to obtain the a, b, c parameters from the numerical load-displacement curve. These values have also been compared with the objective experimental values. Besides, the numerical and experimental load-displacement curves and the stress–strain curves have been compared too. Very good agreement has been observed in all cases. Table 3 gives detail of the comparison between the calculated values $(\sigma_0, n, K)_{calculated}$ and those obtained with standard laboratory traction test $(\sigma_0, n, K)_{known}$. The relative error between the known and calculated values are also shown in Table 3.

Known values		Calculated	
		Value	Error (%)
σ_0	291.6	292.3	0.24
n	0.256	0.2548	0.47
K	854.5	849.76	0.55

Table 3. Results obtained and relative error

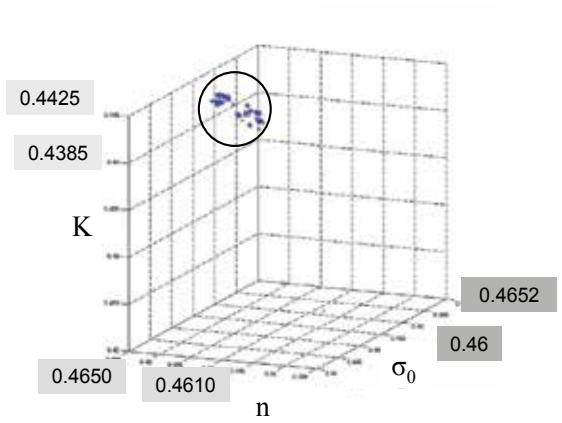


Fig. 12. Pareto front in the space of solutions for another material

Moreover, in Fig. 11 there are different zones with high population density (solutions). Thus, at a slight sought it could be thought that there is no uniqueness in solution, because there are different zones in the figure with high population density. This fact is however observed in some solutions, but generally it is not a problem, because the ranges of variation of variables in the different solutions and their influence on the stress–strain curve is small enough to consider that any result is a good one. However, in many other cases there is only a single zone with high population density and all values trend to a unique solution (Fig. 12)

6.2 Micromechanical characterization

Once the material has been macromechanically characterized, only four of the seven parameters to determine (σ_0, n, K, ε_n, f_n, f_c and f_F) are still unknown (ε_n, f_n, f_c and f_F) and they have to be obtain by means of another inverse procedure. The inputs variables for the micromechanical characterization are ε_n, f_n, f_c and f_F. From Fig. 7 it has been shown that the

only parameters to identify in zone III are ε_n and f_n. In this zone, central composed experiment design centred on faces, based on quadratic response surfaces, has been used to identify the values of these sets of variables to simulate and to choose the minimum number of sets required. In Zone III, only has been applied design of experiments, since defining multiple batteries of simulations for each particular material that it is not known beforehand, is not operative. It has been selected 20 sets of variables (20 experiments) distributed in order to obtain variable inflation factors greater than one and lower than four. In addition, the input variables vary within the following ranges

$$\varepsilon_n = 0.15 - 0.3,\ f_n = 0.01 - 0.07 \tag{14}$$

which are typical ranges for steels (Abendroth and Kuna, 2003). In addition, the maximum variation of (ε_n, f_n) is $\Delta(\varepsilon_n, f_n)_{max}=(0.05, 0.015)$.
Zone III has been adjusted with a linear law in the form $y=l+m\cdot x$. Again, the commercial software DataFit has been used for this purpose. Now, the two output data obtained from each input set are the factors l and m, which depend on the two variables to determine, that is $l=l(\varepsilon_n, f_n)$ and $m=m(\varepsilon_n, f_n)$. Both factors are postulated as second-order polynomial models that can be written in the form

$$g(\varepsilon_n, f_n) = g_0 + g_1 \cdot \varepsilon_n + g_2 \cdot f_n + g_{11} \cdot \varepsilon_n^{\,2} + g_{22} \cdot f_n^{\,2} + g_{12} \cdot \varepsilon_n \cdot f_n \tag{15}$$

where $g(\varepsilon_n, f_n)$ correspond to $l=l(\varepsilon_n, f_n)$ and $m=m(\varepsilon_n, f_n)$.
Coefficients g_{ij} have been obtained using DataFit with regularized input values (ε_n, f_n) varying within the range [−1, 1]. Table 4 gives detail of the values obtained for a 99% confidence interval. Both regressions are very significant and the proportion of variance of l and m, explained are 99.3%, 99.7%, respectively. From zone III of the load-displacement curve of the laboratory small punch test, the target values are $l_{target}=0.0119$ and $m_{target}=0.7909$.

	l	m
g_0	0.022333	0.774075
g_1	-0.001755	0.011094
g_2	0.029950	-0.051640
g_{11}	-0.001519	0.001642
g_{22}	0.000071	0.000086
g_{12}	-0.00123	0.007788

Table 4. g_{ij}. coefficients for the second- order models for functions l and m

Again, Pareto front has been obtained by means of the evolutionary genetic algorithm NSGA-II run in MATLAB. The Pareto front in the space of functions $[(m - m_{target}), (l - l_{target})]$ for the target values is shown in Fig. 13. Moreover, Fig. 14 shows the Pareto front in the space of solutions (ε_n, f_n). In order to verify its 'goodness', the numerical simulation of the resulting values $(\varepsilon_n, f_n)_{calculated}$ has been carried out in order to obtain the (l, m) parameters from the numerical load-displacement curve

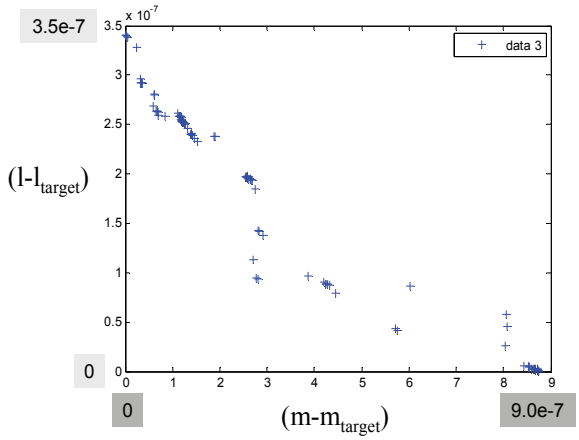


Fig. 13. Pareto front (Zone III) in the space of functions for the target values

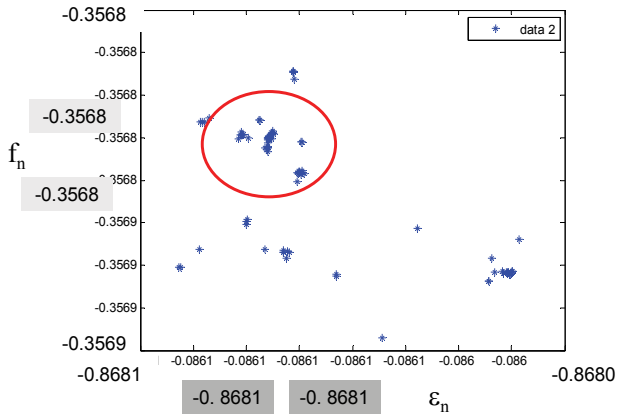


Fig. 14. Pareto front (Zone III) in the space of solutions

Calculated			
ε_n	0.2107		
f_n	0.0293		
↓	Target	Pareto	Error (%)
l	0.0119	0.0119	1.78e-3
m	0.7909	0.7909	1.45e-4

Table 5. Results obtained for zone III and relative error

Table 5 gives detail of these values. This table also shows the relative error of the functions l and m with respect to the objective values $(l,m)_{target}$. Again, very good agreement has been observed in all cases.

Once the parameters ε_n and f_n have been determined a very good agreement between the experimental and numerical curves at zones I, II and III has been achieved. However, it is

from zone IV where the curves separates from each other due to the accelerating effect on the evolution law of the void volume fraction induced from void coalescence, which seriously affect the load resistance capacity of the material. The critical void volume fraction f_c is the only parameter that defines the beginning of coalescence in the material. This value can be obtained from the evolution law of the void volume fraction of the specimen at the region where failure takes place. The value of f_c is the value of porosity (void volume fraction) at the instant in which the experimental and numerical curves begin to separate from each other, and corresponds to the initiation of Zone IV. For the target material (studied material), this separation takes place for a displacement of the punch of 1.32 mm. Thus, the corresponding critical void volume fraction obtained is f_c=0.07 (Fig. 15).
After f_c has been determined, the void volume fraction keeps on growing up to the maximum load point. This maximum marks the beginning of zone V where the load carrying capacity decreases drastically. The slope of this zone depends on f_F. The value of f_F can be obtained carrying out several simulations with different values of f_F until the best agreement in zone V is obtained. For the material studied in this paper (tested by means of the SPT), very good agreement between the experimental and numerical curves is achieved with f_F=0.1.

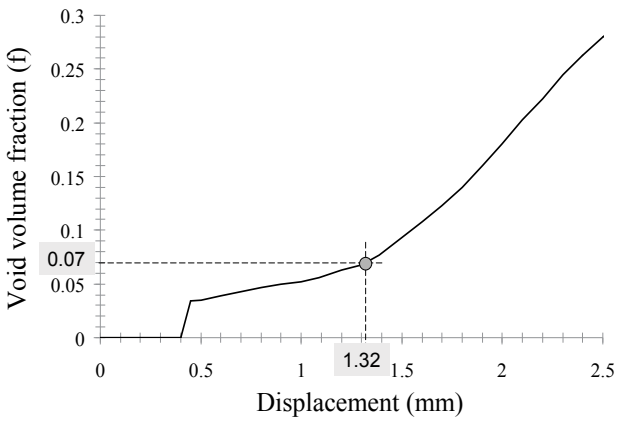


Fig. 15. Void volume fraction–displacement curve and onset of void coalescence

σ_0 (MPa)	n	K	ε_n	f_n	f_c	f_F	q_1	q_2
292.3	0.2548	849.76	0.2107	0.0293	0.07	0.1	1.5	1.0
q_3	f_0	S_n	μ	E (MPa)	ν			
2.25	0	0.01	0.3	200 000	0.3			

Table 6. Complete characterization for the studied material

Once the macromechanic characterization and the micromechanic characterization have been completed, the material is completely characterized. The resulting values for the different parameters obtained by means of the methodology presented in this paper for the complete characterization of the SP tested material are detailed (Table 6).

7. Conclusion

In this paper has been developed an inverse methodology for the determination of the mechanical and damage properties of structural steels that behave according to the Hollomon's law and to the damage model developed by Gurson, Tvergaard and Needleman. Most of these parameters have been derived from the load-displacement curve, which has been obtained by means of small punch tests.

This methodology allows:

1. To characterize not only macromechanically but micromechanically, a wide variety of structural steels, combining experimental data and pseudo-experimental data (numerical simulations).
2. Knowing the deformation of specimen while the test is running
3. To identify the zone of the load-displacement curve that is affected by each variable, and to perform sensitivity analyses.

Moreover, the Pareto front and the evolutionary genetic algorithms allow to obtain, in a relative easy way, numerical results that fit with good agreement the experimental results. In addition, the best way to tackle the parameter identification problem, seems to be the use of a battery of numerical simulations combined with design of experiments. The former has to be used for the macromechanical characterization, whereas the later should be used for the micromechanical characterization.

Finally, the inverse methodology shown in this paper, has to be developed for each type of material, as well as for each thickness of the specimen and each test temperature.

8. References

Abendroth, M. and Kuna, M. (2003) Determination of deformation and failure properties of ductile materials by means of the small punch test and neural networks. *Comput. Mater. Sci*. 28, 633–644, ISSN: 0927-0256.

Bolzon, G. et al. (1997) *Parameter identification of the cohesive crack model Material Identification using Mixed Numerical Experimental Methods*, H. Sol and C. W. J. Oomers Ed., pp. 213–222, Kluver Academic, Dordrecht, Netherlands.

Chu, C. C. and Needleman, A. (1980) Void nucleation effects in biaxially stretched sheets. *J. Eng. Mat. Tech*. 1028, 249–256, ISSN: 0094-4289.

Cuesta, I. I. et al. (2007) Determinación de los parámetros del modelo de daño de Gurson-Tvergaard para la simulación del ensayo de Small Punch. *Anales de Mecánica de la Fractura* 24, 429–434, ISSN: 0213-3725.

DataFit 8.2. Oakdale engineering. Oakdale, California, USA.

Deb, K. (2001). In: *Multiobjective Optimization using Evolutionary Algorithms*. JohnWiley & Sons, Chichester, UK.

Gurson, A. L. (1977) Continuum theory of ductile rupture by void nucleation and growth: part I – yield criteria and flow rules for porous ductile media. *J. Eng. Mat. Tech*. 99, 2–15, ISSN: 0094-4289.

Hibbit, Karlsson and Sorensen (2009) *ABAQUS 6.7*. Inc., Pawtucket, Rhode Island, USA.

Koplik J & Needleman A. (1998)Void growth and coalescence in porous plastic solids. *Int J Solids Struct*;24, 835–53, ISSN: 0020-7683.

Kuehl, R. O. (2000), In: *Design of Experiments*, 2nd edn. Thomson Learning, ISBN: 0-534-36834-4, Duxbury, Massachusetts, USA.

Lucas, G. E.,Odette, G. R., Sokolov,M., Spätig, P.,Yamamoto, T. & Jung, P. (2002). Recent progress in small specimen test technology. *J. Nucl. Mater.* 307–311, 1600–1608, ISSN: 0022-3115.

Lucon, E. (2001) Material damage evaluation and residual life assessment of primary power plant components using specimens of non-standard dimensions. *Mater. Scie. Tech.* 17, 777–785, ISSN: 0861-9786.

MATLAB version 7.3 (R2006b). The MathWorks.

Montgomery, D. O. (1997) *Design and Analysis of Experiments, 4th edn*. John Wiley and Sons, New York.

Myers, R. H. & Montgomery, D. O. (1995) *Response Surface Methodology*, JohnWiley and Sons Ed., ISBN: 0-471-41255-4 New York.

Peñuelas, I., Rodríguez, C., Antuña, M., Betegón, C. & Lezcano, R. (2009). Caracterización mecánica de recubrimientos mediante ensayos miniatura de punzonamiento, *Actas del IX congreso iberoamericano de Ingeniería Mecánica* , sec 13, 25-31, ISBN:. 978-84-692-8516-9.

Rodriguez, C. et al (2009) Mechanical Properties Characterization of Heat-Affected Zone Using the Small Punch Test, *Welding journal*, 88, 9, 188-192, ISSN: 0043-2296.

Seshadri, A. (2006). A fast elitist multiobjective genetic algorithm: NSGA-II. *MATLAB Central*.

Stravroulakis, G. E., Bolzon, G. & Waszczyszyn, L. (2003). Inverse analysis. *Comprehensive Struct. Integrity*. 3, 1–34,.

Sun DZ, et al., (1992) Application of micro-mechanical models to the prediction of ductile fracture. *Fracture mechanics, 22nd symposium.* ASTM STP 1131, vol. II, 368–78.

Tvergaard, V. (1981) Influence of voids on shear bands instabilities under plane strain conditions. *Int. J. Fract.* 17, 389–407, ISSN: 0376-9429.

Tvergaard, V. (1982) On localization in ductile materials containing spherical voids. *Int. J. Fract*. 18, 157–169, ISSN: 0376-9429.

Tvergaard, V. & Needleman, A. (1984). Analysis of cup-cone fracture in a round tensile bar. *Acta Metall.* 32, 57–169, ISSN: 0-56-7151.

Tvergaard, V. (1990) Material failure by void growth to coalescence. *Adv. Appl. Mech.* 27, 83–151 ISBN: 0-12-002040.

14

Laser Shock Peening: Modeling, Simulations, and Applications

Y.B. Guo
Dept. of Mechanical Engineering,
The University of Alabama,
Tuscaloosa, AL 35487
U.S.A.

1. Introduction

Laser shock peening (LSP) is a surface treatment process to improve surface integrity and fabricate micro surface structures. The mechanism of LSP is shown in Figure 1. LSP is a cold mechanical process where pressure waves caused by expanding plasma plastically deform the surface of a material. LSP uses a thin layer of ablative material that is opaque to the laser. The opaque ablative material, typically black spray paint or tape, is used as a sacrificial layer in the early study by Fairland and Clauer (Fairland & Clauer, 1976). The sacrificial layer also minimizes undesirable thermal effects on the surface caused by the laser. The laser partially vaporizes the ablative layer to form high pressure plasma. The plasma, confined by a thin layer of water film, expands rapidly resulting in a recoiling pressure wave on the order of GPa reported by Fairland et al. (Fairland et al., 1972), Fabbro et al. (Fabbro et al., 1990), Masse and Barreau (Masse & Barreau, 1995), Berthe et al. (Berthe et al., 1997), Fan et al. (Fan et al., 2005), Warren, et al. (Warren et al., 2008), and Caslaru, et al. (Caslaru et al., 2008). The pressure wave is the cold mechanical process that plastically deforms the surface. The plasma-induced shock pressure on the order of GPa can be much larger than the dynamic yield strength of the work material. Once the peak pressure exceeds material yield strength, the transient shock pressure causes severe plastic deformation, refined grain size, compressive residual stresses, and increased hardness at the surface and in the subsurface. As a result, the mechanical properties on the workpiece surface are enhanced to improve the performance of fatigue, wear, corrosion and foreign object damage.

Besides producing favorable surface integrity, LSP can also be used to fabricate various micro surface structures such as dent arrays using an automatic x-y positioning system. The micro surface structures may have various functions. For example a laser peened dent array can act as lubricant reservoirs to reduce coefficient of friction in bearings and to reduce flow drag of compressor blades.

Just due to the transient nature of shocking pressure, real time in-situ measurement of laser/material interaction is very challenging. A numerical simulation method may provide an ideal tool to shed light on the process mechanics and resultant surface integrity.

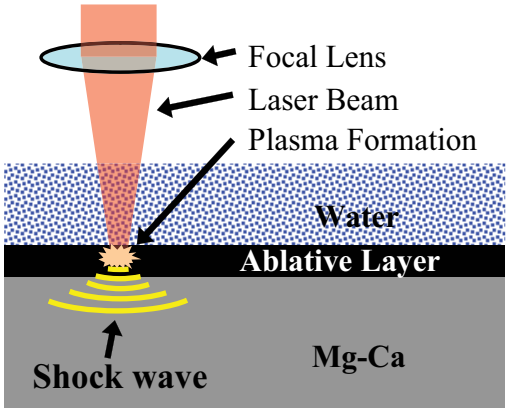


Fig. 1. Process principle of micro dent fabrication by LSP

2. State-of-the-art of LSP simulation

2.1 LSP for enhanced surface integrity

LSP is a surface treatment process to modify surface properties for improved wear and fatigue performance. LSP is primarily conducted on metallic components. The principle of LSP is to use a high intensity laser and suitable overlays to generate high pressure shock waves on the workpiece surface.

An increase in fatigue strength is achieved by large magnitudes of compressive residual stresses which develop in the subsurface. The maximum compressive residual stress is usually on the surface and decreases with depth. The transient shock waves can also induce microstructure changes near the surface and cause high density of dislocations. The combined effect of the microstructure changes and dislocation entanglement contribute to improved surface properties.

It has been shown by previous research (Clauer et al., 1983; Clauer & Koucky, 1991; Peyre et al., 1996; Vaccari, 1992; Ashley, 1998; Brown, 1998; Banas et al., 1990) that improved fatigue life of metallic components such as bearings, gears, shafts, etc can be accomplished by inducing compressive residual stress using LSP. An advantage of LSP is that the affected depth is very deep (≈ 1 mm) as compared with other surface treatment processes such as conventional shot peening.

During LSP (Figure 1), the sample surface is first coated with a thin layer of material such as black paint which is opaque to the laser beam. This opaque layer acts as sacrificial material and is converted to high pressure plasma as it absorbs energy from a high intensity laser (1-10 GW/cm^2) for very short time durations (< 100 ns). If the sample surface is also submerged in a transparent media such as water, the rapidly expanding plasma cannot escape and the resulting shock wave is transmitted into the sample subsurface. The shock pressure can be much larger than the dynamic yield strength of the material (>1 GPa), which causes surface plastic deformation and compressive residual stresses which can extend to a deep depth (≈ 1 mm) in the subsurface. Due to the high strains/strain rates that the material experiences, there can also be significant microstructure changes thus causing the surface properties such as hardness, strength, and fatigue strength to be improved. Because thermal rise in the sample is nearly eliminated by the water overlay, LSP is primarily a cold working process.

A significant amount of LSP research has been conducted to investigate the surface integrity. Most experimental work has focused on the determination of residual stress magnitudes and distributions in the near surface. The effect of LSP on surface properties and fatigue life has been relatively less studied. The resulting surface integrity can be correlated with the LSP process parameters such as laser intensity, laser spot size, peening pass, and peening spacing. The following is a brief overview of previous research results.
Residual stress can vary with LSP process parameters. Increasing the laser intensity increases both the magnitude and affected depth of compressive stress in the subsurface. However, it has been shown that laser intensities greater than a particular threshold serve to decrease the surface stress magnitude, but continue to increase the magnitude and affected depth in the subsurface (Peyre et al., 1996). This was attributed to expansion release waves that are formed due to high energy shock waves. An investigation of laser spot size effect showed that energy attenuation is less for larger spot sizes allowing the stress shock wave to propagate deeper into the material (Fabbro et al., 1998). Thus larger spot sizes increase the depth of plastic deformation. A study of overlapped laser spots (Clauer & Koucky, 1991; Peyre et al., 1996; Peyre et al., 1998; Ruschau et al., 1999) showed that the residual stress distribution is nearly uniform and is entirely compressive.
Previous numerical simulations of LSP have been performed to gain better understanding of the physical process. Because LSP is a highly transient process, it is difficult (if not impossible) to experimentally observe and quantify the stress wave propagation into the sample surface. Simulations have been used to aid in determining accurate shock pressure models, verify experimental data, and predict residual stress profiles. Zhang et al. (Zhang et al., 2004) improved the shock pressure models by Clauer (Clauer & Holbrock, 1981) and Fabbro (Fabbro et al., 1990) by accounting for the non-linear mass transfer of LSP. The model also accounts for the time dependent radial expansion of plasma for micro sized laser peening. Finite element simulations have been performed to verify and predict residual stress profiles after LSP (Braisted & Brockman, 1999; Ding, 2003; Zhang & Yao, 2002).

2.2 LSP fabrication of micro dent arrays

The controlled patterning of solid surfaces improves the wear, friction and lubrication (Anderson et al., 2007). Micro dents serve as fluid reservoirs that effectively retain lubricant. Also micro dents function as traps for wear debris, eliminating a potential plowing effect caused by entrapped particles. The long term benefit of surface patterning is to extend the life of contacting surfaces. Micro dents on the surface can improve the surface lifetime by a factor of ten (Romano et al., 2003). Experimental studies on the effect of dent patterns on micro-grooved sapphire discs lead to the conclusion that fabricated micro dents on metallic surfaces is a useful method to reduce friction in sliding contact. Manufacturing techniques to fabricate micro dents arrays on component surfaces include micro indentation (Nakatsuji & Mori, 2001), micro-drilling (Friedrich, 2002), and laser ablation (Etsion, 2005). These processes often induce surface damage such as cracks and phase transformation which may shorten component life. A new process to make dents while avoid material damage is highly needed. When the pressure exceeds the dynamic yield stress in LSP, plastic deformation occurs and forms a dent on the surface. LSP is a flexible and economic technique to fabricate micro dent arrays on metallic component surfaces using an automatic x-y positioning system.

2.3 LSP biomaterials

Biodegradable implants are a relatively new and emerging form of treatment for common bone ailments. Biodegradable implants are useful to the healing process due to the ability to gradually dissolve and absorb into the human body after implantation. The development of biodegradable implants has had a beneficial effect on in-vivo treatment of patients with various bone ailments.

Currently, biodegradable implants are mainly made of polymers, such as poly-L-Lactic acid. However, these polymer based implants usually have an unsatisfactory mechanical strength. An alternative to biodegradable polymer implants is permanent metallic implants composed of steel or titanium alloys. Permanent metal implants have superior strength compared to polymers. As a consequence, metal implants are often too stiff resulting in a stress shielding effect that can be damaging to the healing process (Benli et al., 2008; Completo et al., 2008; Au et al., 2007; Shi et al., 2007; Isaksson & Lerner, 2003; Nagels et al., 2003; Gefen, 2002). Stress shielding occurs when bone is shielded by an implant from carrying load. As a result, the bone tends to weaken over time resulting in more damage. To minimize the effects of stress shielding on the human body while still retaining strength, a soft lightweight metal is required. Therefore, Mg alloys are proposed as an ideal biodegradable implant material due to its biocompatibility and superior strength to weight ratio compared to that of other biomaterials.

Magnesium is an element essential to the human body. Intake of a certain amount of magnesium (300 ~ 400 mg/day) is normally required for regular metabolic activities (Seiler, 1987). The direct corrosion product of magnesium, Mg^{2+}, is easily absorbed or consumed by the human body (Song, 2007). However, the rapidly generated by-products of magnesium corrosion, such as hydrogen gas and hydroxides, are not physiologically favorable. Hydrogen evolution and alkalinization resulting from corrosion of Mg are the most critical obstacles in using magnesium as an implant material. A straightforward strategy to tackle these difficulties is to control the corrosion rate of a biodegradable magnesium implant. The adjustment of surface property is one promising solution to control the corrosion rate of Mg in human body.

In this chapter, calcium (Ca) was alloyed with Mg to form a Mg-Ca alloy. It is well known that Ca is a major component in human bone and is also essential in chemical signaling with cells (Ilich & Kerstetter, 2000). Ca has a low density (1.55 g/cm^3) such that when alloyed with Mg, the density is similar to that of bone. The Ca in Mg-Ca alloys produces hydroxyapatite (HA) as a corrosion product on the surface of the implant. HA mineral is a naturally occurring form of calcium apatite with the formula $Ca_{10}(PO_4)_6(OH)_2$ and has close resemblance to the chemical and mineral components of teeth and bone. As a result of this similarity it stimulates bone cells to attack the implant surface and make proper bonding (Aksakal & Hanyaloglu, 2008), which allows for fractured segments to realign in correct anatomical position which is critical to recovery.

Laser shock peening (LSP) is a promising surface treatment technique to improve the surface integrity by imparting compressive residual stresses that are beneficial for controlling corrosion of Mg-Ca implants. LSP has been initiated to fabricate an array of dents on component surfaces (Warren et al., 2005; Warren & Guo, 2007; Caslaru et al., 2008; Sealy & Guo, 2008). Previous finite element analyses (FEA) of LSP investigate individual peening of a metal substrate. FEA of single peens neglects the effect of neighboring dents on topography, hardness and residual stress. The purpose of this chapter is to determine the

effects of sequential peening of Mg-Ca alloy on surface topography as well as predict the residual stress profile. Sequential peening experiments and simulations were performed and compared to single peening experiments and simulations.

3. LSP modeling and simulation procedures

3.1 Modeling of 3D spatial and temporal shock pressure

Because the laser spot is circular, a two-dimensional finite element simulation can not reflect the true nature of LSP. For this reason a 3D model must be used for realistic simulation of the laser induced shock pressure. The simulation mesh is shown in Figure 2. The mesh has two regions with different mesh densities. With a high mesh density, the results from a simulation converge to a unique solution. As expected, the area where the pressure is applied contains a higher mesh density than the outer regions of the model. The dense mesh region consists of elements of 1 μm cubes. Micron elements provide a suitable spatial resolution of the output variables.

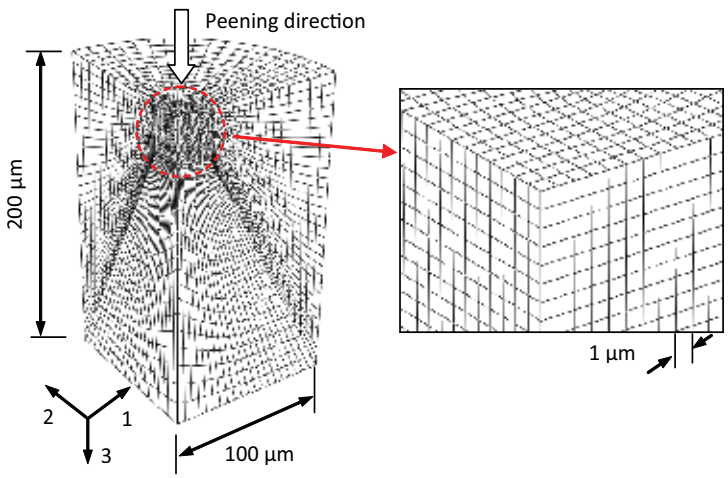


Fig. 2. Three-dimensional (3D) FEA simulation of LSP

The spatial and temporal pressure distribution during LSP is neither uniform nor linear. For this reason a subroutine VDLOAD was used to apply the non-uniform shock pressure. The subroutine allows the pressure intensity to vary simultaneously with respect to radial distance from the center of the laser spot and elapsed time of the laser pulse. It works by assigning local origins at the center of the desired shock peen locations and calculates the radial distance to each node surrounding this new origin from the equation of a circle as

$$r = \sqrt{\left(curcoord(i,1)\right)^2 + \left(curcoord(i,2)\right)^2} \tag{1}$$

where $curcoord(i,1)$ and $curcoord(i,2)$ are the coordinates in the 1 and 2 directions, respectively, for the current node at each time increment of the analysis.

The pressure as a function of radial distance from the center of the laser spot follows a Gaussian distribution (Zhang et al., 2004). Maximum pressure is located at the center of the laser spot and decreases with increasing radial distance from the center.

The pressure distribution is also a function of the elapsed time of laser pulse. The pressure is initially zero and reaches a peak value when the elapsed time equals the total pulse time. Following the results by Zhang, et al. (Zhang et al., 2004), the pressure versus time can be well represented as fourth order polynomials to follow the pressure vs. time relationships shown in Figure 3.

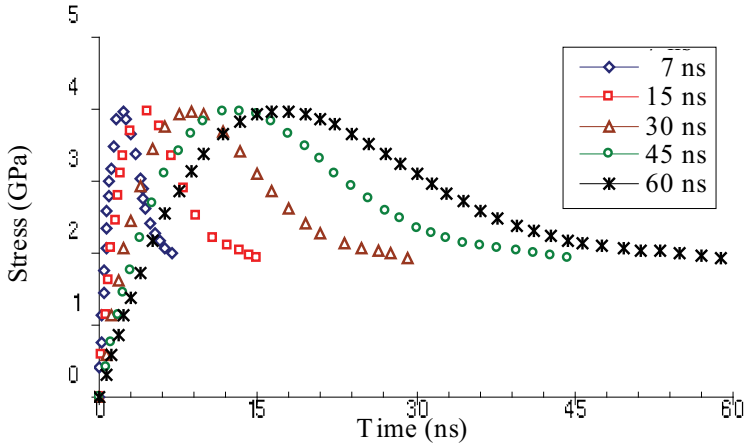


Fig. 3. Theoretical pressure vs. time curve

The pressure $P(r,t)$ at any point and time can be calculated as

$$P(r,t) = P(t)\exp\left(-\frac{r^2}{2R^2}\right) \tag{2}$$

where $P(t)$ is the pressure at time t during the laser pulse interpolated from Figure 3, r is the radial distance from the center of the laser spot in Eq. (1), and R is the laser spot radius.

3.2 Modeling of dynamic mechanical behavior

Due to the extremely high strain rates (> 10^6 s^{-1}) that occur during LSP, traditional material models are not adequate. For this reason a subroutine VUMAT was used to incorporate the plasticity/failure model developed by the internal state variable (ISV) plasticity model (Bammann et al., 1993; Bammann et al., 1996). The BCJ constitutive equations can be written below.

$$\overset{\circ}{\underline{\boldsymbol{\sigma}}} = \dot{\underline{\boldsymbol{\sigma}}} - \underline{\mathbf{W}}^{\mathbf{e}}\underline{\boldsymbol{\sigma}} + \underline{\boldsymbol{\sigma}}\underline{\mathbf{W}}^{\mathbf{e}} = \lambda tr(\underline{\mathbf{D}}^{\mathbf{e}})\underline{\mathbf{I}} + 2\mu\underline{\mathbf{D}}^{\mathbf{e}} \tag{3}$$

$$\underline{\mathbf{D}}^{\mathbf{e}} = \underline{\mathbf{D}} - \underline{\mathbf{D}}^{\mathbf{p}} \tag{4}$$

$$\underline{\mathbf{D}}^{\mathbf{p}} = f(T)\sinh\left[\frac{\left\|\underline{\boldsymbol{\sigma}} - \underline{\boldsymbol{\alpha}}\right\| - \{R + Y(T)\}}{V(T)}\right]\frac{\underline{\boldsymbol{\sigma}} - \underline{\boldsymbol{\alpha}}}{\left\|\underline{\boldsymbol{\sigma}} - \underline{\boldsymbol{\alpha}}\right\|} \tag{5}$$

$$\begin{aligned}\overset{\circ}{\underline{\boldsymbol{\alpha}}} &= \dot{\underline{\boldsymbol{\alpha}}} - \underline{\mathbf{W}^{\mathbf{e}}}\,\underline{\boldsymbol{\alpha}} + \underline{\boldsymbol{\alpha}}\,\underline{\mathbf{W}^{\mathbf{e}}} \\ &= h(T)\underline{\mathbf{D}^{\mathbf{p}}} - \left[\sqrt{\frac{2}{3}}r_d(T)\left\|\underline{\mathbf{D}^{\mathbf{p}}}\right\| + r_s(T)\right]\|\underline{\boldsymbol{\alpha}}\|\underline{\boldsymbol{\alpha}}\end{aligned} \tag{6}$$

$$\dot{R} = H(T)\underline{\mathbf{D}^{\mathbf{p}}} - \left[\sqrt{\frac{2}{3}}R_d(T)\left\|\underline{\mathbf{D}^{\mathbf{p}}}\right\| + R_s(T)\right]R^2 \tag{7}$$

The evolution equations (6) and (7) for the internal state variables $\underline{\alpha}$ and R are motivated from dislocation mechanics and are in a hardening-minus-recovery format. The kinematic hardening internal state variable $\underline{\alpha}$ representing directional hardening is related to the dislocations in cell interior. The variable captures the softening effect due to unloading, also termed as Bauschinger's effect. The isotropic hardening internal state variable R is related to the dislocations in walls and it captures the continued hardening at large strains. The use of internal state variables and the evolution equations enable the prediction of strain rate history and temperature history effects.

The model uses nine temperature dependent functions to describe the inelastic response. They can be classified into three basic types: those associated with the initial yield, the hardening functions, and the recovery functions. The rate-independent yield stress $Y(T)$, the rate-dependence of initial yield stress $f(T)$, and the magnitude of rate-dependence of yield stress $V(T)$ are assumed to be of the forms

$$V(T) = C_1 \exp(-C_2 / T) \tag{8}$$

$$Y(T) = C_3 \exp(C_4 / T)([1 + (\tanh(C_{19}(C_{20} - T)))] / 2) \tag{9}$$

$$f(T) = C_5 \exp(-C_6 / T) \tag{10}$$

The three functions of $r_d(T)$, $h(T)$, $r_s(T)$ describe the tensor or kinematic hardening and recovery, which can be thought of as the center of yield surface. The functions of $R_d(T)$, $H(T)$, and $R_s(T)$ describe the scalar or isotropic hardening and recovery, which can be thought of as the radius of the yield surface.

$$r_d(T) = C_7 \exp(-C_8 / T) \tag{11}$$

$$h(T) = C_9 - C_{10}T \tag{12}$$

$$r_s(T) = C_{11} \exp(-C_{12} / T) \tag{13}$$

$$R_d(T) = C_{13} \exp(-C_{14} / T) \tag{14}$$

$$H(T) = C_{15} - C_{16}T \tag{15}$$

$$R_s(T) = C_{17} \exp(-C_{18} / T) \tag{16}$$

The material constants ($C_1 - C_{20}$) can be determined by fitting the BCJ model to the baseline test data using a non-linear square fitting method. The very short pulse duration (< 100 ns) makes the simulation an ideal transient case. For this purpose, Abaqus/Explicit (HKS, 2008) was used to implement the simulation scheme.

4. Simulation case studies

3D finite element simulation models in peening several engineering materials have been developed to investigate transient laser/material interactions at nano timescale during peening. Three application case studies in automotive, aerospace, and biomedical industries are presented using the developed simulation method.

4.1 Case 1: LSP simulation of enhancing surface integrity of hardened steel

The purpose of this case study is to micro laser shock peening hardened AISI 52100 steel (62 HRc) by varying the laser pulse duration (time elapsed for maximum pressure) for times of 5, 10, 50, and 100 ns. For comparative purposes, a conventional material model which uses experimental compression stress/strain data and the failure/plasticity model termed the ISV model is be used to predict the material behavior. The results will provide insight into the highly transient LSP process and assist in proper selection of experimental parameters for control of surface integrity requirements after LSP.

The fitted material constants are shown in Table 1. The simulation was performed as a single pass of laser shock peening with a laser spot radius of 6 μm. The simulated laser intensity is 5.5 GW/cm² which attains a maximum pressure of ≈ 4 GPa. The laser pulse time was varied as 5, 10, 50, and 100 ns in order to test the effect of strain rate on the transient stress and strain.

BCJ Parameter	Material Constants	BCJ Parameter	Material Constants
C1 (MPa)	1.00E+00	C11 (s/MPa)	2.39E-03
C2 (K)	1.00E+00	C12 (K)	4.00E+02
C3 (MPa)	2.52E+03	C13 (1/MPa)	5.00E-02
C4 (K)	5.85E+01	C14 (K)	0.00E+00
C5 (1/s)	1.00E+00	C15 (MPa)	1.50E+02
C6 (K)	-1.20E+04	C16 (MPa/K)	-1.40E+01
C7 (1/MPa)	4.00E-02	C17 (s/MPa)	2.70E-03
C8 (K)	0.00E+00	C18 (K)	0.00E+00
C9 (MPa)	5.60E+03	C19	4.15E-03
C10 (MPa/K)	9.00E+00	C20 (K)	6.65E+02

Table 1. ISV material constants of AISI 52100 steel

The greatest magnitude (stress or strain) during the simulation was retrieved across and beneath the laser spot as shown in Figure 4. This allows direct comparison of various laser pulse times on the transient behavior of the material during LSP. For comparative purposes, the results are plotted for simulations using the BCJ model and direct data input in table format (hereafter "Table") which use only compression stress/strain data for modeling material behavior.

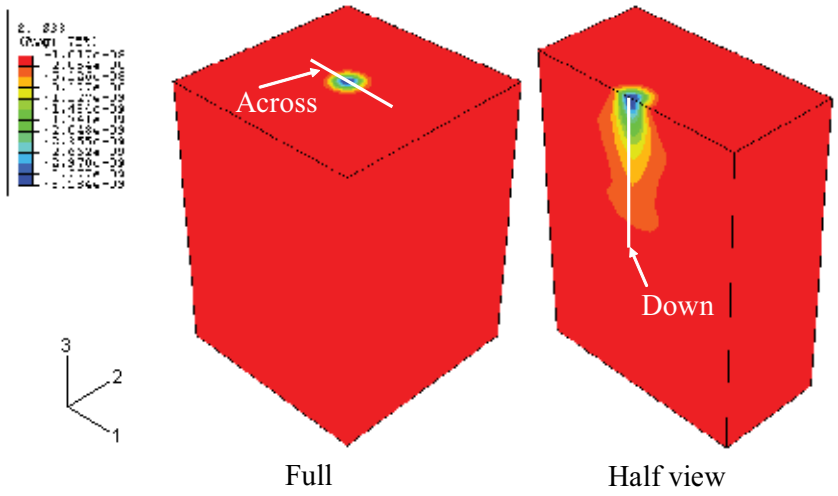


Fig. 4. Result path locations

4.1.1 Stress distributions

The maximum subsurface normal stress in the peening direction is shown in Figure 5a. The maximum stress occurs on the surface for the greater pulse times (50 ns and 100 ns) while it occurs in the subsurface (≈ 3.5 μm) for the lower pulse times (5 ns and 10 ns). This may be due to higher strain rates generated by the shorter pulse times. However, the stress at all depths greater than 3.5 μm is more compressive for the shorter pulse durations. It is observed that the subsurface stress difference at the same depth can be as much as 750 MPa between the shortest and longest laser pulse times. Another observation is the consistently higher stress (at depths > 3.5 μm) predicted by the BCJ model than that for simulations using table format. This is reasonable due to the extremely high strain rates during LSP for which there is no experimental data available. At pulse times of 50 ns and 100 ns, the strain rate has less influence and the stress distribution curves are nearly identical for the BCJ model and table format.

The maximum normal stress across the specimen surface is shown in Figure 5b. From the figure it is observed that the difference between the experienced surface stress at the laser center can be as large as 1.0 GPa by varying the laser pulse time. However, the difference is negligible beyond the diameter of the laser spot (12 μm).

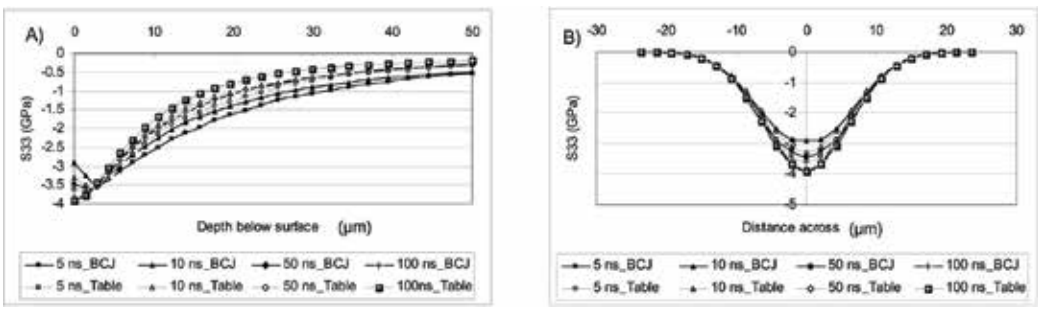


Fig. 5. Stress variation (peening direction a) down and b) across

The subsurface von Mises profile is shown in Figure 6a. The maximum value of von Mises stress occurs at a depth of 4.2 μm for all simulation cases. It is also observed that the stress magnitude is inversely proportional to the laser pulse time. The difference between the 5 ns and 10 ns pulse times is, however, much larger (500 MPa) than for the 50 ns and 100 ns cases (50 MPa) at the surface showing that the relationship is not linear. In addition, the variation of the stress for the 5 ns and 10 ns pulse is larger than that for the 50 ns and 100 ns pulse times when comparing the BCJ model and table format.
Figure 6b shows the von Mises distribution across the top surface. The trend is similar to that of the transverse normal stress in that the largest magnitude occurs across the entire surface by order of decreasing pulse time. A sharp rise in von Mises stress occurs across a diameter of ≈ 24 μm reaching a maximum at the center of the laser spot. The influence of the high strain rate induced by the 5 ns pulse is seen by the 30% higher equivalent stress when compared to the next pulse time (10ns).

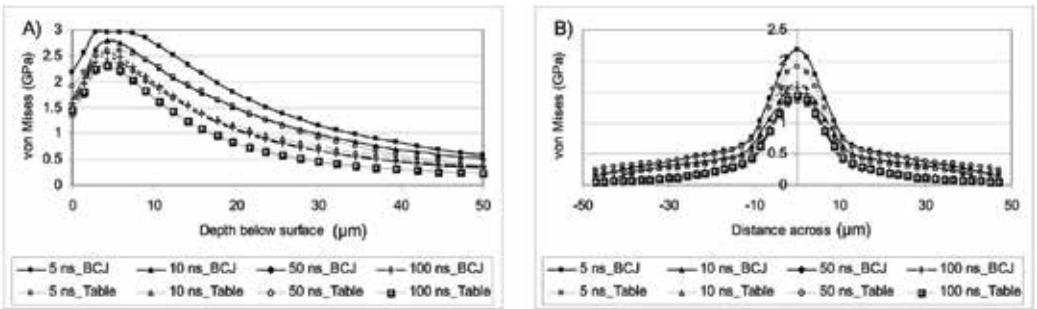


Fig. 6. Von mises variation a) down and b) across

4.1.2 Strain rate

The maximum strain occurred in the loading direction and is shown in Figure 7a. For each case, the greatest strain magnitude occurred in the subsurface, the depth of which is dependent on the pulse duration. For the 10, 50, and 100 ns cases, the maximum value occurred at a depth of ≈ 2.8 μm, while the 5 ns case reached a maximum strain of -1.87×10^{-2} at a depth of 4.3 μm. After the maximum strain is reached, the strain magnitude decreases with the highest value occurring at each depth in order of decreasing pulse duration.

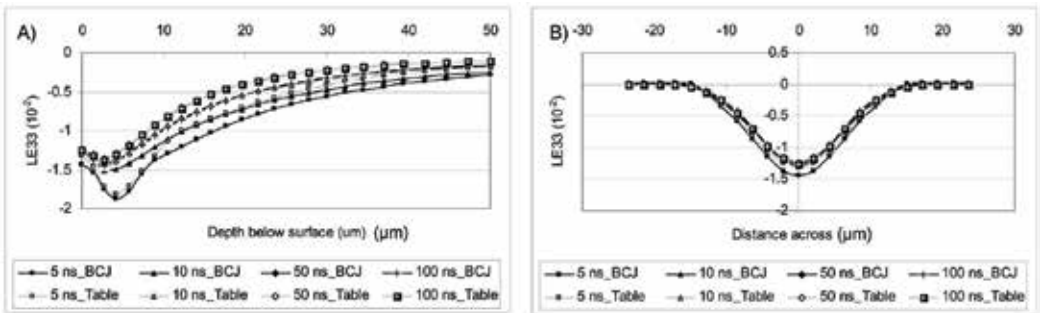


Fig. 7. Strain variation a) down and b) across

The maximum strain in the loading direction across the surface is shown in Figure 7b. The maximum value of -1.4×10^{-2} was attained for the 5 ns pulse time using the BCJ model. A ≈ 7% lower strain was predicted by the table format for each simulation pulse time. The maximum strain attained by the 10, 50, and 100 ns cases was ≈ -1.2×10^{-2}.

4.1.3 Residual stress

The predicted residual stresses were obtained from the surface element located at the center of the laser spot. A comparison of the measured and simulated residual stress values are shown in Figure 8. Both the predicted and measured residual stresses are compressive, so they agree with the nature and trend. There is some discrepancy between the two which may be due to several factors that differentiate the experimental procedure from the simulation. In addition to numerical errors, the first is the massive parallel LSP used for the experiment which was not accounted for in the benchmark simulation. The overlaps of consecutive laser peenings that occurred in LSP experiments would increase the magnitudes of compressive residual stress. The predicted residual stresses from both single and two LSP passes are expected to be lower than those from the experiments. The second is that the x-ray diffraction technique using $Cr_{K\alpha}$ radiation actually measures an average residual stress in the depth of x-ray penetration (5-10 μm). In addition, the exact location of residual stress measurement with regard to the laser peened zone can not be accurately controlled for the experiment. For the measurement itself, the residual stress magnitudes across the peened surface are different just due to the nonuniform nature of surface integrity. Unless high precision calibration and control can be carried out first, the x-ray and other non-destructive measurement methods are only useful for comparative purpose.

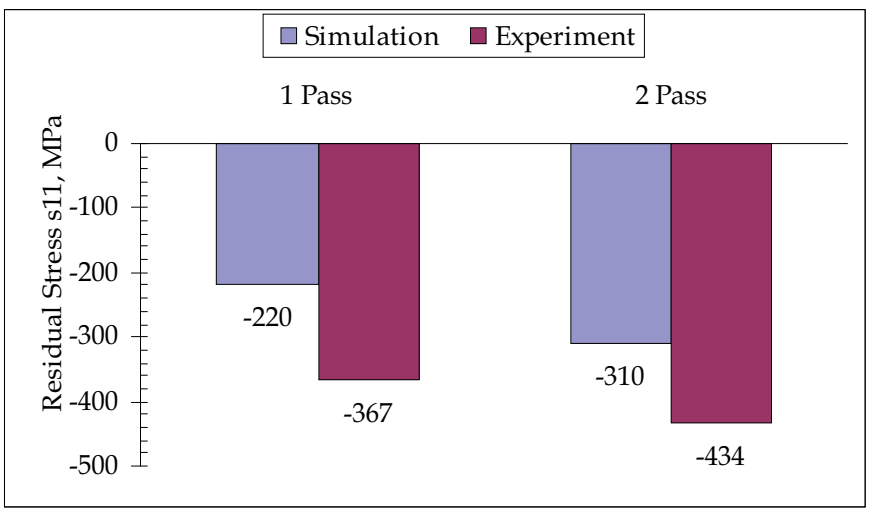


Fig. 8. Comparison of predicted surface residual stress s11 with measured data

4.2 Case 2: LSP simulation of fabricating micro dent arrays on titanium surface

A 3D finite element simulation model was developed to fabricate micro dent arrays on titanium Ti-6Al-4V surfaces as shown in Figure 2, for improving tribology performance. Ti-6Al-4V is a widely used engineering material in aerospace, automotive, and biomedical

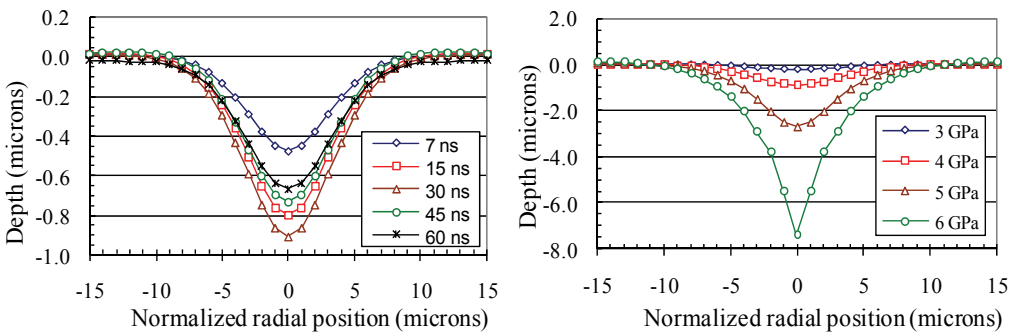
industries. Micro surface structures of the LSP processed Ti-6Al-4V components is critical for product performance. However, the surface deformation and mechanical behavior in patterning a Ti-6Al-4V surface has not been well understood. The simulation aims to understand the laser/material interaction and the related mechanical phenomena. The material constants (C_1 – C_{20}) were determined by fitting the ISV model to the baseline test data using a non-linear square fitting method. The fitted material constants are shown in Table 2 (Guo et al., 2005). The modulus of elasticity for Ti-6Al-4V is 114 GPa. Poisson's ratio is 0.34 at room temperature. The density is 4430 kg/m^3.

ISV parameter	Material constants	ISV parameter	Material constants
C1 (MPa)	1.0	C11 (s/MPa)	205
C2 (K)	0.2	C12 (K)	0
C3 (MPa)	1570	C13 (1/MPa)	1.9E-3
C4 (K)	10	C14 (K)	0
C5 (1/s)	1.0E-5	C15 (MPa)	619
C6 (K)	0	C16 (MPa/K)	3.8E-1
C7 (1/MPa)	7.0E-2	C17 (s/MPa)	5.0E-4
C8 (K)	0	C18 (K)	0
C9 (MPa)	1866	C19	1.0992E-3
C10 (MPa/K)	0.3	C20 (K)	876

Table 2. ISV material constants of Ti-6Al-4V

4.2.1 Simulated dent geometry

Figure 9a depicts the dent profiles for the various pulse times. Each dent was measured 50 ns after the simulation. Initially, increasing the pulse time leads to an increase in depth. However, the 30 ns simulation has the maximum depth at 0.9 μm. The simulations with pulse times greater than 30 ns exhibited a decrease in the depth. This suggests there is an optimal pulse time which produces the deepest dents given a peak pressure.



(a) constant peak pressure of 4 GPa (b) constant pulse time of 30 ns

Fig. 9. Simulated dent profiles

Figure 9b shows the dent profiles as the peak pressure increases. There is a non-linear relationship between the dent depth and peak pressure. As the load increases, the depth of the dent increases as well. However, the radius of each dent is about 20 microns. A comparison between the simulated dent contours and measured ones will be conducted in a future study.

4.2.2 Surface material behavior at different peening time

Material behaviors at the surface are characterized by the stress/strain graphs along the peening or depth direction (axis-3). Each stress/strain profile plotted represents the maximum transient stress/strain during the peening process. The corresponding radial curves are corresponding stress/strain graphs where the maximum occurs along the depth.

Transient stress profiles: Von Mises stress along the depth is plotted in Figure 10a. In each simulation, the maximum von Mises is 1.45 GPa and occurs about 3 μm below the surface and gradually decreases to 1.27 GPa. The stress then sharply decreases toward zero as the depth increases. Surface material at different peening times experiences similar von Mises characteristics but at different depths. Figure 10b shows von Mises profile in the radial direction 3 μm in the subsurface. In the radius of 9 μm, the von Mises stress remains greater than 1.2 GPa. Then, the stress begins to decrease exponentially.

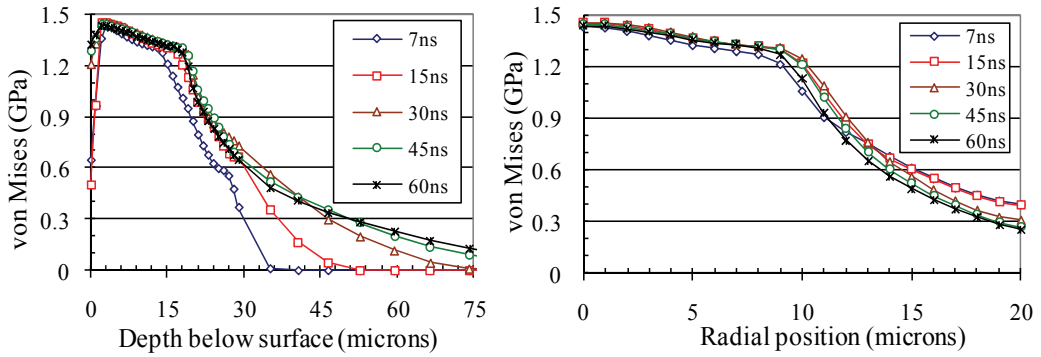


Fig. 10. von Mises stress distributions at different peening time

Transient strain profiles: The effective plastic strain PEEQ along the depth, Figure 11a, exhibits an inverse relationship with the peening time. The plastic strain decreases with the increased peening time. However, below the surface that is not the case. The 30 ns peening time induces the maximum plastic strain. PEEQ converges to zero at 15 μm to 20 μm in subsurface. Figure 11b illustrates the radial profiles of PEEQ which extends 10 μm in the radial direction.

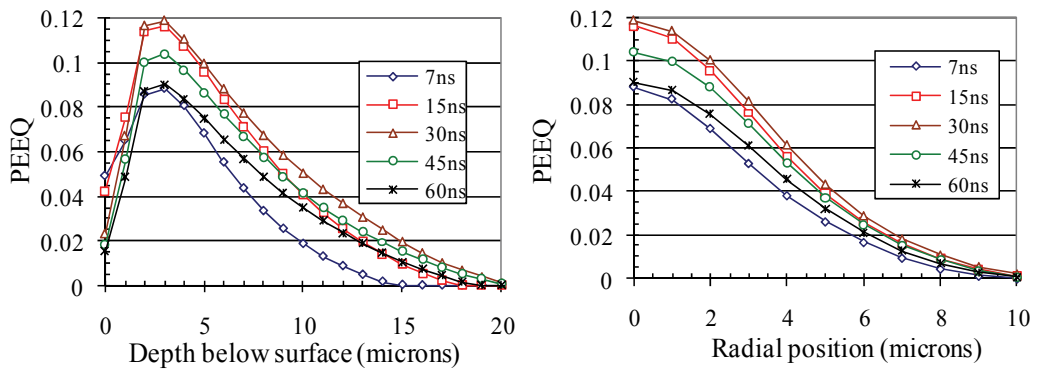


Fig. 11. Effective plastic strain distributions at different peening time

Strain rate profiles: Figure 12a shows the strain rate along the depth for each peening time. Material at the 7 ns peening case experiences the largest strain rate at 31×10^6/s at 3 μm in the subsurface. As peening time increases, the strain rate decreases non-linearly. In each case, the peak rate occurs at 2 to 3 μm below the surface. Figure 12b shows the radial profiles of the strain rate which extends approximately 10 μm from the peening center. The strain rate for the 7 ns case converges more rapidly in the radial direction than other cases.

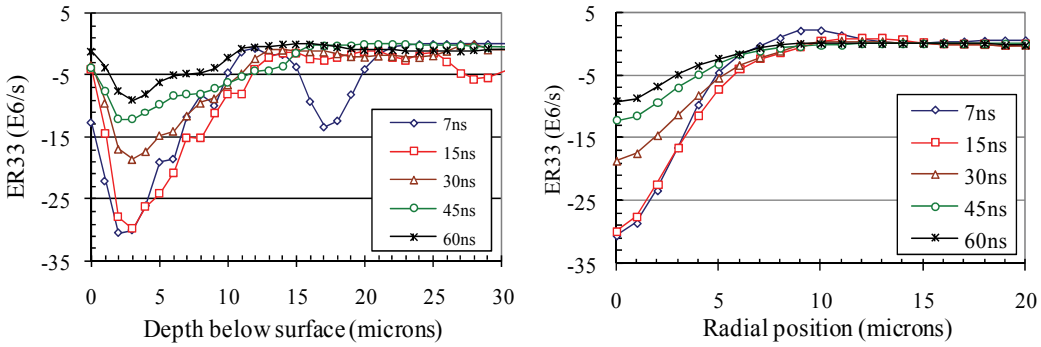


Fig. 12. Strain rate distributions at different peening time

4.2.3 Surface material behavior at different peening pressure

Transient stress profiles: Von Mises profiles in the depth are plotted in Figure 13a. At peak pressures 3 GPa and 4 GPa, the maximum von Mises occurs at 3 μm in the subsurface. As peak load increases the maximum von Mises moves toward the surface. It is also observed that von Mises profiles overlap at peak pressures 5 GPa and 6 GPa. It implies that increasing the peak pressure over 6 GPa will saturate von Mises stress. Initially, the stress gradually decreases along the depth. Once it decreases to 1.3 GPa, it rapidly drops and converges toward zero. Figure 13b shows the stress along the radial direction. It exhibits a similar phenomenon seen in the depth direction.

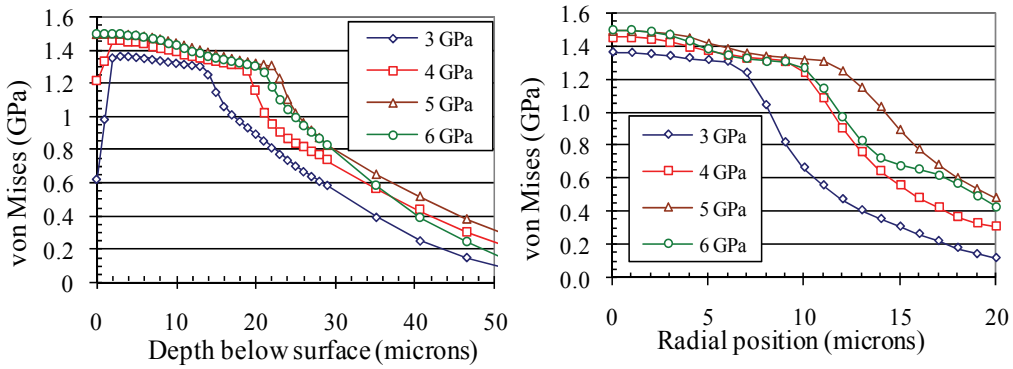


Fig. 13. von Mises stress distributions at different peening pressure

Strain profiles: The equivalent plastic strain in the depth is plotted in Figure 14a. The maximum plastic strain at 6 GPa peak pressure is on the surface, while it moves deeper into the subsurface as the peak load decreases. For example, it moves to 3 μm deeper for the case of 3 GPa peak pressure. The corresponding radial profiles for PEEQ are shown in Figure 14b.

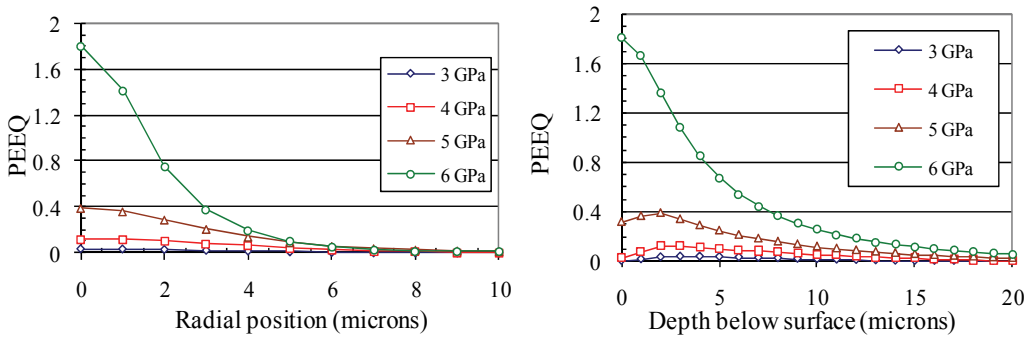


Fig. 14. Effective plastic strain distributions at different peening pressure

Strain rate profiles: Figure 15a shows that the maximum strain rate is 226×10^6/s on the surface at 6 GPa peak pressure. As peak pressure decreases, the maximum strain rate moves deeper below the surface. In addition, the simulations at peak pressures of 3 GPa and 4 GPa experienced much smaller strain rates ($< 2\times10^6$) on the surface. But the maximum strain rates occur at 3 μm in the subsurface. The corresponding radial profiles of the strain rate in Figure 15b extend approximately 6 μm from the peening center.

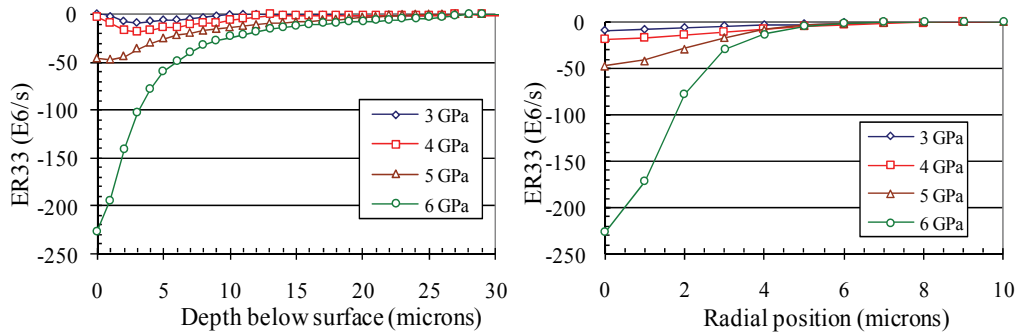


Fig. 15. Strain rate distributions at different peening pressure

4.3 Case 3: LSP simulation of peening biomedical material for enhanced corrosion performance

A 3D semi-infinite model was used to simulate micro scale laser shock peening of biodegradable Mg-Ca. The material constants (C_1 – C_{20}) of the biomaterial were determined by fitting the ISV model to the baseline test data using a non-linear square fitting method (Guo et al., 2005). The fitted material constants are shown in Table 3 (Guo & Salahshoor, 2010). The modulus of elasticity for Mg-Ca is 45 GPa. Poisson's ratio is 0.33 at room temperature. The density is 1750 kg/m^3.

A series of four simulations were performed in order to simulate sequential LSP. The Mg-Ca surface was peened once per simulation. Each simulation is composed of two steps. In the first step, the shock pressure is applied on the top surface. Next, the stresses and strains are allowed sufficient time to relax so that the solution has time to stabilize. The results from the first simulation were imported to the second simulation and so on until the surface was peened 4 times.

ISV parameter	Material constants	ISV parameter	Material constants
C_1 (MPa)	1.0	C_{11} (s/MPa)	1E-4
C_2 (K)	600	C_{12} (K)	0
C_3 (MPa)	850	C_{13} (1/MPa)	0.7
C_4 (K)	20	C_{14} (K)	100
C_5 (1/s)	1.0E-7	C_{15} (MPa)	3E4
C_6 (K)	0	C_{16} (MPa/K)	39
C_7 (1/MPa)	0.1	C_{17} (s/MPa)	380
C_8 (K)	-300	C_{18} (K)	-900
C_9 (MPa)	2500	C_{19}	0.2
C_{10} (MPa/K)	0	C_{20} (K)	312.8

Table 3. ISV material constants of Mg-Ca alloy

4.3.1 Simulation scheme

The 3D model in Figure 16 contains a quarter cylinder of 70,818 C3D8R finite elements and 3,575 CIN3D8 infinite elements. The quarter cylinder mesh allows for a comprehensive analysis of the three dimensional stress and strain behavior below the surface while minimizing the computation time. Infinite elements as quiet boundary along the back and bottom surfaces were implemented to allow for stress waves to pass through a non-reflective boundary.

The mesh has two regions with different mesh densities. As expected, the area where the pressure is applied contains a higher mesh density than the outer regions of the model. The dense mesh region consists of 30 µm wide cubic elements. Micron level elements provide a suitable spatial resolution of the output variables to ensure spatial convergence.

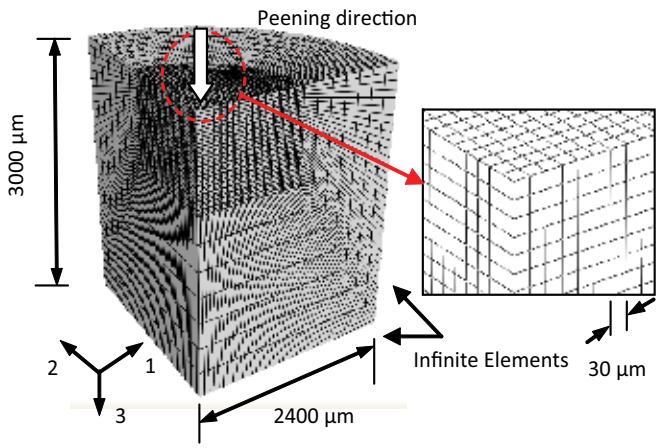


Fig. 16. Three-dimensional FEA simulation of LSP

The pressure induced by LSP is a function of elapsed time and radial position. A useful approximation for *P(t)* is to assume it follows a 6th order polynomial as shown in Figure 17. The generic profile is based on numerous researchers (Berthe et al., 1997; Fabbro et al., 1990; Devaux et al., 1993; Wu & Shin, 2005; Zhang et al., 2004) who have measured the *P(t)* as a function of time. The critical components of *P(t)* are the pulse time and the peak pressure. The pressure pulse time typically last 2-3 times longer than the laser pulse (Devaux et al., 1993; Berthe et al., 1999; Zhang & Yao, 2002). For the purpose of these simulations, the pressure pulse was assumed to be 3 times longer than the 7 ns laser pulse. The peak pressure for *P(t)* in water confined regime was estimated by

$$P(GPa) = 0.01\sqrt{\frac{\alpha}{2\alpha+3}}\sqrt{Z(g/cm^2s)}\sqrt{I_o(GW/cm^2)} \tag{17}$$

where P is the peak pressure, Z is combined shock impedance defined by the following Eq. (18), I_o is the power density given by Eq. (19), and a is a correction factor for the efficiency of the interaction (Fabbro et al., 1990; Peyre et al., 1996). Since the ablative material used in these experiments was relatively thick and absorbent compared to other materials used in literature, a was estimated to be low (0.1) such that the majority of the energy was absorbed by the ablative material. Z_{MgCa} is defined as the product of the density and shock velocity ($Z_{MgCa}=\rho_{MgCa}U_{MgCa}$). The density of Mg-Ca is 1750 kg/m^3 and the shock velocity is approximated based on the wave speed of sound through Mg-Ca (≈ 5000 m/s). Z_{MgCa} and Z_{water} are 8.75x10^5 and 1.65x10^5 g/cm^2, respectively.

$$\frac{2}{Z} = \frac{1}{Z_{MgCa}} + \frac{1}{Z_{water}} \tag{18}$$

$$I_o = \frac{E}{t_p A} \tag{19}$$

where E is the average energy per pulse given as 0.2667 J. t_p is the simulated pressure pulse time (21ns). A is the cross-sectional area of the generated plasma. The diameter of the pressure wave is approximately 250 μm which results in a peak pressure of 5 GPa.

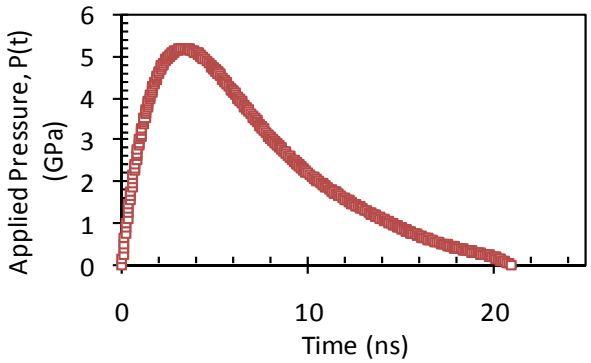


Fig. 17. Spatially uniform shock pressure, *P(t)*

In this study, the radial expansion of plasma was taken into account for the following reasons. First, the experimental laser spot size is on the order of 100 microns. With such a small spot size, the expansion of plasma may not be neglected in the radial direction. Furthermore, the experimental ablative layer is not fully vaporized because it is thick and absorbs energy well. As a consequence, the pressure wave generated by the plasma has time and space to expand in all directions before entering the metal substrate. Radial expansion of plasma was modeled by allowing the applied pressure to act perpendicular to the deformed surface. Initially the pressure is one dimensional. As deformation occurs, the pressure follows the deformed surface resulting in a spherical shape pressure that expands in the radial direction.

Implementing the temporal and spatial shock pressure is very challenging and a user load subroutine is therefore required. The user subroutine VDLOAD (Warren et al., 2008) of shock pressure has been programmed to apply a non-uniform shock pressure across the top surface. The circular pressure was applied in four locations. Figure 18 shows the peening distribution along the top surface. The spacing between simulated peens is 800 μm.

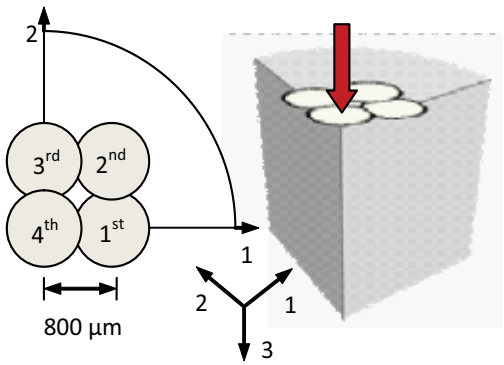


Fig. 18. Sequence of peening simulations (quarter shown)

4.3.2 Simulation results

Material behavior is characterized by stress/strain graphs along the peening or depth direction (axis-3 in Figure 18) and radial directions (axis-1&2). Each stress/strain profile represents the stabilized residual stress/strain. Residual stress/strain was achieved 30 μs after the pressure pulse.

Dent geometry: Figure 19a depicts the simulated dent profiles for sequential and single LSP. The diameter of the simulated dents was 600-700 μm and had a depth of 10 μm. There was a negligible effect of neighboring dents on the overall dent depth. However, it was observed that neighboring dents do influence the tensile pile up region. The magnitude of the pile up increased approximately 50%. It is believed to be due to the radial expansion of neighboring peens. Tensile pile up is critical to tribological applications such as implants. A tensile region on the surface can drastically affect the wear and fatigue performance of a surface. Figure 19b shows the experimental dent profiles for sequential and single LSP. The experimental dents also had a diameter between 600-700 μm and a depth of 11μm. Results from the experiments confirms the validity of the simulation. Figure 19c and 19d are optical images of dents by sequential and single LSP.

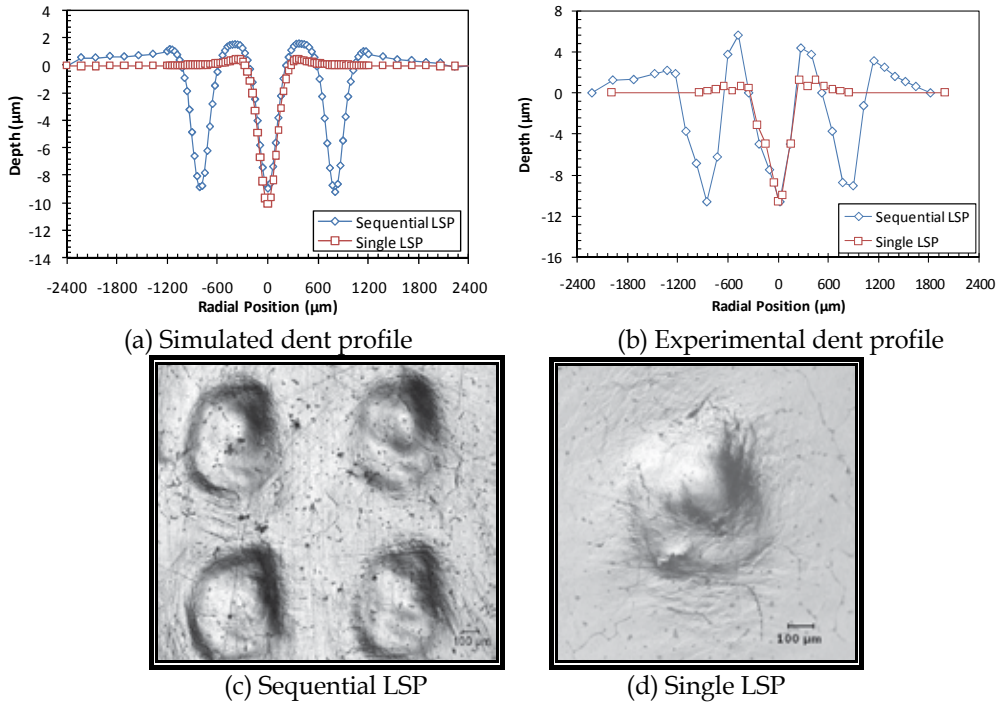


Fig. 19. Simulated and measured dent topography

Residual stress profiles: The predicted residual von Mises stress and S33 stress along the depth direction are shown in Figures 20a and 21a. The von Mises stress penetrated deeper into the surface for sequential peening. As expected, sequential peening had a greater effect on the surface residual stress since a larger area was exposed to peening. Along the depth direction, the residual stress S33 is compressive for approximately 150 µm. The compressed region is followed by a tensile region that eventually approaches 0 MPa. The magnitude of the compressive residual stress below the surface is 23 MPa. The predicted residual von Mises stress and S22 stress along the radial direction are shown in Figures 20b and 21b. Single peening neglects the effects from neighboring stress fields on the surface residual stress. Future work will include comparing simulated residual stress profiles to experimental residual stress.

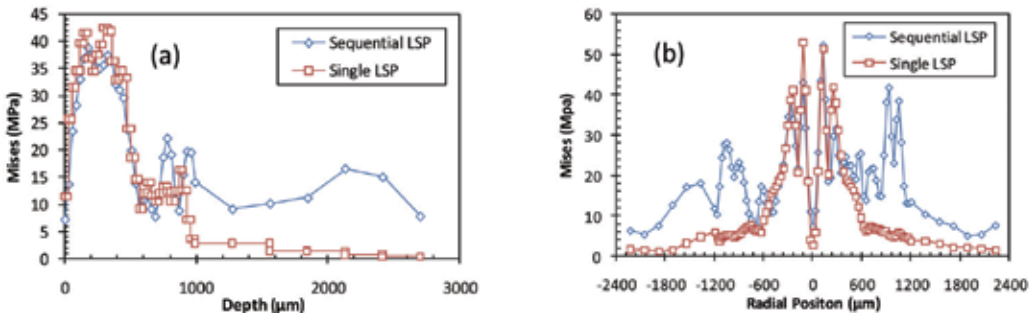


Fig. 20. Residual von Mises stress along depth (a) and radial (b) directions

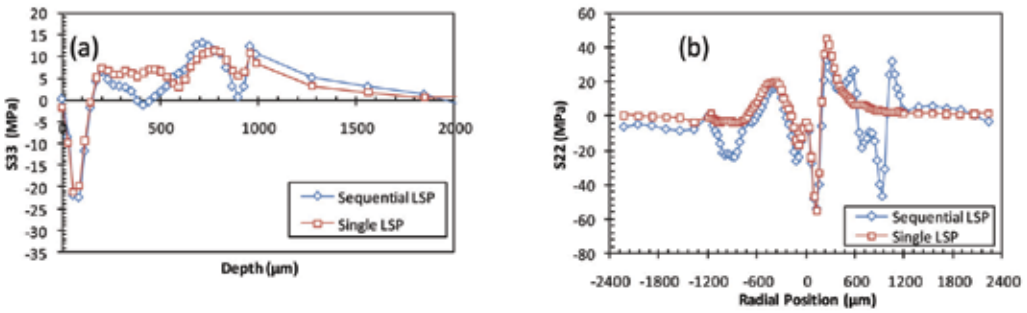


Fig. 21. Residual stress, S33 along depth (a) and S22 along the radial (b) directions

Strain and strain rate profiles: The plastic strain in the depth and radial directions is shown in Figure 22. The plastic strain extended 500 μm below the surface. The residual stress from previous peens had a negligible effect on the plastic strain. The maximum plastic strain occurred on the top surface and in the center of the dent. The diameter of the plastic zone is directly related to the topography of the dent. The peak strain rate in peening direction for the simulations was 19×10^6 s^{-1} in Figure 23.

This work focuses on the experiment and FEA simulation of LSP MgCa alloy. More experimental results are needed to verify the simulation results. Further work is needed to demonstrate the effectiveness of the resulting surface by this method in improving surgery of bone ailments.

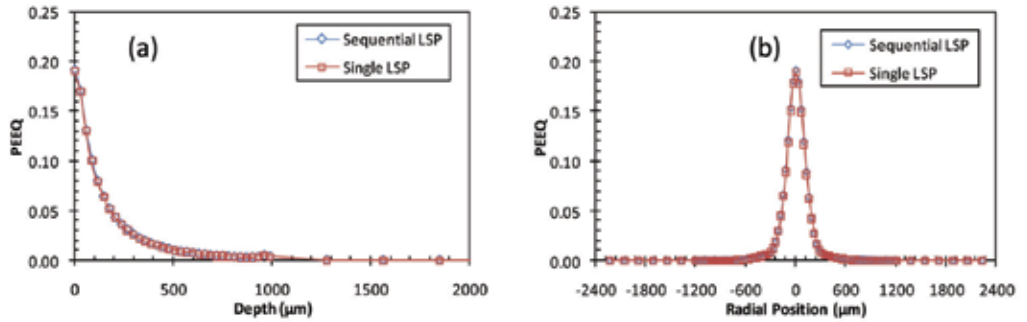


Fig. 22. Equivalent plastic strain PEEQ along depth (a) and radial (b) directions

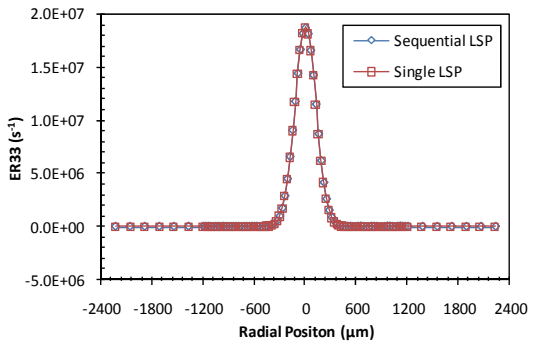


Fig. 23. Strain rate on the top surface, ER33

5. Conclusions

Laser shock peening (LSP) is a surface treatment process to improve surface integrity which significantly impacts component performance of fatigue, wear, corrosion, and foreign object damage. This chapter provides a state-of-the-art of LSP simulation and discussed the challenging issues to simulate a LSP process using finite element method. The new contributions of this chapter provide a 3D model of temporal and spatial shock pressure and material user subroutine of dynamic mechanical behavior at high strain rates. Three simulation case studies in automotive, aerospace, and biomedical industries are presented using the developed simulation method. The key results may be summarized as follows.

- The 3D spatial and temporal peening pressure was modeled using a user subroutine.
- The dynamic material behavior at high strain rates was modeled using the ISV model. Material constants of three types of important engineering materials were obtained.
- The simulated dent geometry and residual stresses are similar to the measured data. This suggests the pressure model used successfully characterized the formation and propagation of the pressure wave.
- The results suggested there is an optimal peening time that produces the deepest dent. Pulse time has a significant effect on the strain rate range.
- The maximum transient stress occurred at a certain peening time. The stress along the radial direction was slightly affected by the peening times. However, the stress along the depth and radius were drastically affected by the peek pressures. Increasing the peak pressure resulted in larger and shallower maximum stress.
- Sequential peening affects the dent topography by increasing the size of the tensile pile up region. The pile-up region forms from the radial expansion of plasma. It is believed to have a great significance on tribological aspects of the biodegradable implant material.
- There was no observed effect on the depth of dents when sequential peening was used as opposed to individual peening.

6. References

[1] Fairland, B.P. & Clauer, A.H. (1976). Effect of water and paint coatings on the magnitude of laser-generated shock waves. *Optics Communications*, 14(3), 588-591.

[2] Fairland, B.P.; Wilcox, B.A.; Gallagher, W.J. & Williams, D.N. (1972). Laser shock-induced microstructural and mechanical property changes in 7075 aluminum. *J. Appl. Phys.*, 43, 3893-3895.

[3] Fabbro, R.; Fournier, J.; Ballard, P.; Devaux, D. & Virmont, J. (1990). Physical study of laser-produced plasma in confined geometry. *J. App. Physics*, 68, 775-54.

[4] Masse, J.E. & Barreau, G. (1995). Laser generation of stress waves in metal. *Surf. Coatings Tech.*, 70, 179-191.

[5] Berthe, L.; Fabbro, R.; Peyre, P.; Tollier, L. & Bartnicki, E. (1997). Shock waves from a water-confined laser-generated plasma. *J. Appl. Phys.*, 82, 2826-2832.

[6] Fan, Y.; Wang, Y.; Vukelic, S. & Yao, Y.L. (2005). Wave-solid interactions in laser-shocked-induced deformation processes. *J. Appl. Phys.*, 98 (10), 104904-104901-11.

[7] Warren, A.W.; Guo, Y.B. & Chen, S.C. (2008). Massive parallel micro laser shock peening: simulation, validation, and analysis. *Int. J. Fatigue*, 30, 188-197.

[8] Caslaru, R.; Sealy, M.P. & Guo, Y.B. (2009). Fabrication and characterization of micro dent array on al 6061-t6 surface by laser shock peening. *Trans. NAMRI/SME*, 37, 159-166.

[9] Clauer, A.H.; Ford, C.T. & Ford, S.C. (1983). The effects of laser shock processing on the fatigue properties of T-3 aluminum, In: *Lasers in materials processing,* American Society for Metals, 7-22, Metals Park.

[10] Clauer, A.H. & Koucky, J.R. (1991). Laser shock processing increases the fatigue life of metal parts. *Materials and Processing*, 6, 3-5.

[11] Peyre, P.; Fabbro, R.; Merrien, P. & Lieurade, H.P. (1996). Laser shock processing of aluminum alloys. Application to high cycle fatigue behavior, *Materials Science and Engineering A,* 210, 102-113.

[12] Vaccari, J.A. (1992). Laser shocking extends fatigue life. *American Machinist*, 62-64.

[13] Ashley, S. (1998). Powerful laser means better peening. *Mechanical Engineering*, 120, 12.

[14] Brown, A.S. (1998). A shocking way to strengthen metal, *Aerospace America*, 21-23.

[15] Banas, G.; Elsayed-Ali, H.E.; Lawrence, F.V. & Rigsbee, J.M. (1990). Laser shock-induced mechanical and microstructural modification of welded maraging steel. *Journal of Applied Physics,* 67, 2380-2384.

[16] Fabbro, R.; Peyre, P.; Berthe, L. & Sherpereel, X. (1998). Physics and application of laser-shock processing. *Journal of Laser Applications*, 10, 265-279.

[17] Peyre, P.; Berthe, L.; Scherpereel, X. & Fabbro, R. (1998). Laser-shock processing of aluminum coated 55C1 steel in water-confinement regime, characterization and application to high-cycle fatigue behavior. *Journal of Materials Science*, 33, 1421-1429.

[18] Ruschau, J.J.; John, R.; Thompson, S.R. & Nicholas, T. (1999). Fatigue crack nucleation and growth rate behavior of laser shock peened titanium. *International Journal of Fatigue*, 21, 199-209.

[19] Zhang, W.; Yao, Y.L. & Noyan, I.C. (2004). Microscale laser shock peening of thin films, Part 1: Experiment modeling and simulation. *Journal of Manufacturing Science and Engineering*, 126, 10-17.

[20] Clauer, A.H. & Holbrock, J.H. (1981). Effects of laser induced shock waves on metals, *Proceedings of Shock Waves and High Strain Phenomena in Metals-Concepts and Applications*, pp. 675-702, Plenum, New York.

[21] Braisted, W. & Brockman, R. (1999). Finite element simulation of laser shock peening. *International Journal of Fatigue*, 21, 719-724.

[22] Ding, K. & Ye, L. (2003). Three-dimensional dynamic finite element analysis of multiple laser shock peening process. *Surface Engineering*, 19, 351-358.

[23] Zhang, W. & Yao, Y.L. (2002). Micro scale laser shock processing of metallic components. *Journal of Manufacturing Science and Engineering*, 124, 369-378.

[24] Anderson, P.; Koskinen, J.; Varjus, S.; Gerbig, Y.; Haefke, H.; Georgiou, S.; Zmhud, B. & Buss, W. (2007). Microlubrication effect by laser-textured steel surfaces, *Wear*, 262, 369-379.

[25] Romano, V.; Weber, H.P.; Dumitru, G.; Pimenov, S.; Kononenko, T.V.; Konov, V.; Haefke, H. & Gerbig, G. (2003). Laser surface microstructuring to improve tribological systems. *Proceedings of the SPIE*, 5121, 199-211.

[26] Nakatsuji, T. & Mori, A. (2001). The Tribological Effect of Electrolytically Produced Micro-pools and Phosphoric Compounds on Medium Carbon Steel Surfaces in Rolling-Sliding Contact. *Tribology Transactions*, 44, 173-178.

[27] Friedrich, C.R. (2002). Micromechanical machining of high aspect ratio prototypes. *Microsystem technologies*, 8, 343-347.
[28] Etsion, I. (2005). State of Art in Laser Surface Texturing. *Journal of Tribology*, 127, 248-253.
[29] Benli, S.; Aksoy, S.; Havitcioglu, H. & Kucuk, M. (2008). Evaluation of bone plate with low stiffness material in terms of stress distribution. *Journal of Biomechanics,* 41, 3229-3235.
[30] Completo, A.; Fonseca, F. & Simoes, J.A. (2008). Strain shielding in proximal tibia of stemmed knee prosthesis: experimental study. *Journal of Biomechanics*, 41, 560-566.
[31] Au, A.G.; Raso, V.J.; Liggins, A.B. & Amirfazli, A. (2007). Contribution of loading conditions and material properties to stress shielding near the tibial component of total knee replacements. *Journal of Biomechanics*, 40, 1410-1416.
[32] Shi, J.F.; Wang, C.J.; Laoui, T.; Hart, W. & Hall, R. (2007). A dynamic model of simulating stress distribution in the distal femur after total knee replacement, *Proceedings of the Inst MECH E Part H, Journal of Engineering in Medicine*, 221, 903-912.
[33] Isaksson, H. & Lerner, A.L. (2003). Mathematical modeling of stress shielding with bioresorbable materials for internal fracture fixation, *Proceedings of Bioengineering Conference*, 1041-1042, Key Biscayne, Florida.
[34] Nagels, J.; Stokdijk, M. & Rozing, P. M. (2003). Stress shielding and bone resorption in shoulder arthroplasty. *Journal of Shoulder and Elbow Surgery*, 12, 35-39.
[35] Gefen, A. (2002). Computational simulations of stress shielding and bone resorption around existing and computer-designed orthopedic screws. *Medical & Biological Engineering & Computing*, 40, 311- 322.
[36] Seiler, H. G. (1987). *Handbook on Toxicity of Inorganic Compounds*, CRC Press.
[37] Song, G. (2007). Control of biodegradation of biocompatible magnesium alloys. *Corrosion Science*, 49, 1696-1701.
[38] Ilich, J. Z. & Kerstetter, J. E. (2000). Nutrition in bone health revisited: a story beyond calcium. *Journal of the American College of Nutrition*, 19, 715-737.
[39] Aksakal., B. & Hanyaloglu, C. (2008). Bioceramic dip-coating on Ti-6Al-4V and 316L SS implant materials. *Journal of Materials Science: Materials in Medicine*, 19, 2097-2104.
[40] Warren, A.W.; Guo, Y.B. & Chen, S.C. (2005). A numerical simulation of massive parallel laser shock peening, *Proc. of ASME International Mechanical Engineering Congress & Exposition*, Orlando, FL.
[41] Warren, A.W. & Guo, Y.B. (2007). FEA modeling and analysis of 3d pressure and mechanical behavior at high strain rate in micro laser peening. *Trans. NAMRI/SME*, 35, 409-416.
[42] Sealy, M.P. & Guo, Y.B. (2008). Fabrication and finite element simulation of μ-laser shock peening for micro dents. *Int. J. Comp. Methods in Eng. Sci. & Mech.*, 10, 149-157.
[43] Bammann, D.J.; Chiesa, M.L.; Horstemeyer, M.F. & Weingarten, L.I. (1993). Failure in ductile materials using finite element methods, In: *Structural Crashworthiness and Failure,* Jones, N. & Weirzbicki, T. (Eds.), 1–54, Elsevier, 1851669698, Amsterdam.
[44] Bammann, D.J.; Chiesa, M.L. & Johnson, G.C. (1996). Modeling large deformation and failure in manufacturing processes, In: *19th International Congress on Theoretical and Applied Mechanics,* Tatsumi, T., Watanabe, E. & Kambe, T., (Eds.), 359–376, Elsevier, Amsterdam.

[45] HKS, Inc. (2008). *ABAQUS User's Manual, Ver. 6.4*, Pawtucket, RI .

[46] Guo, Y.B.; Wen, Q. & Horstemeyer, M.F. (2005). An Internal State Variable Plasticity Based Approach to Determine Dynamic Loading History Effects in Manufacturing Processes. *Int. J. Mech. Sci.*, 47, 1423-1441.

[47] Guo, Y.B. & Salahshoor, M. (2010). Process Mechanics and Surface Integrity by High-Speed Dry Milling of Biodegradable Magnesium-Calcium Implant Alloys. *Ann. CIRP*, 59/1, 151-154.

[48] Devaux, D.; Fabbro, R.; Tollier, L. & Bartnicki, E. (1993). Generation of shock waves by laser-induced plasma in confined geometry. *Journal of Applied Physics*, 74, 2268-2273.

[49] Wu, B. & Shin, Y. (2005). A self-closed thermal model for laser shock peening under the water confinement regime configuration and comparisons to experiments. *Journal of Applied Physics*, 97, 1-12.

[50] Berthe, L.; Fabbro, R.; Peyre, P. & Bartniki, E. (1999). Wavelength dependent laser shock-wave generation in the water-confinement regime. *Journal of Applied Physics,* 85-11, 7552-7555.

15

Numerical and Physical Simulation of Pulsed Arc Welding with Forced Short-Circuiting of the Arc Gap

Oksana Shpigunova and Anatoly Glazunov
Institute of Applied Mathematics and Mechanics of Tomsk State University
Russia

1. Introduction

The essence and complicacy of approach to computer design of an optimal pulsed arc welding technology is that programmed periodic action should be developed on the one hand to exert its effect on melting and transfer of an electrode metal, and on the other hand, to control over the molten pool fluidity, the structural formation of weld and heat-affected zone (HAZ) that appears as result of the weld pool crystallization whilst ensuring stability of the pulsed regime in welding in different spatial positions. The results of physical simulation and mathematical modelling permit to design optimal algorithms of pulsed control of energy parameters of welding - arc current and voltage, arc heated efficiency, peak short-circuiting current. The results of computer experiments permit to establish pulsed welding controlled parameters - service properties of welded joints (such as the sizes of welds and HAZ, quality and strength properties of welded joints) relation.

The solution of the pulsed arc welding and surfacing processes optimizing problem is a matter of great significance because of continuously increasing requirements on quality and reliability of welded joints, saving in welding fabrication cost. The construction of welded structures has a number of special features. These are associated with the character of welding metallurgy and solidification processes in the weld metal, the welded joint heating and cooling conditions and others, influenced on the stability of parameters of the complex electrodynamics' system: "power source – electrode – arc – weld pool – welded joint". It is necessary to ensure the regulation of the penetration depth, welding in wider gaps and in different spatial positions, joining metals and alloys of dissimilar chemical composition, decreasing the degree of splashing of electrode metal, increasing the stability of arc ignition and arcing. Arc heating sources energy concentration is unable to solve these technological problems including increasing the productivity of welding operations and improving the welded joints quality parameters.

The rate of assembling operations in the construction of pipelines is increased mainly as a result of automation of welding non-rotating joints. The main part of the system of transmission pipelines in Russia for the transport of natural gas, oil and products of processing mainly of the high-pressure type and with a large diameter (1220-1420 mm) has been operating for a relatively long period of time: 30% of gas pipelines have been operating for more than 20 years and 15% for more than approximately 30 years. In order to maintain

the pipelines in good operating condition, it is necessary to carry out either running repairs of defective areas with the application of effective and universal technologies, or replace defective sections completely in individual long areas. In addition, the expansion of the existing network of transmission pipelines, used for the transport of oil and gas to neighbouring countries, requires the application of more productive methods of welding and technological means for the realization of these methods.

The advantages of new high-productivity methods of mechanized welding in CO_2 and gas mixtures and also with self-shielding flux-cored wires include the decrease in the welding time of the root and filling layers, decreases in the dimensions of the cross-section of the welding gap and, correspondingly, in the volume of deposited metal, and increases in the productivity of welding and assembly operations. However, the mechanized welding methods also have disadvantages, associated with the presence of defects in welded joints, lack of fusion at the edges and between the layers, determined by the instability of the welding process, continuous changes of the spatial position of the welding pool and more extensive splashing of electrode metal.

Further progress in welding fabrication ensuring higher rate of assembling and repair, lowering of the welding operations cost, while providing the required level of quality and service properties of the welded joints, is possible by development of new high-efficient adaptive pulsed welding technologies and specialized equipment for their implementation.

In contrast to the well-known methods of arc welding, including pulsed methods, using "rigid" control programmes, the adaptive pulsed processes are based on the correction of selected algorithm through feedback channels on the basis of instantaneous values of the main energy parameters of the welding process in relation to the condition of the "power source $\rightarrow$ arc $\rightarrow$ weld pool $\rightarrow$ welded joint zone" control object.

Such parameters as: the arc voltage; duration of typical stages of microcycle - arcing time in the pulse, the break with the duration t_p; instantaneous and mean values of current; arc power in a separate microcycle; melting energy of every electrode metal droplet can be the main controlled parameters of adaptive pulsed technological process.

The adaptive pulsed technological process of welding in comparison with the stationary one permits:

- to control the processes of melting and droplet transfer of electrode metal, the solidification in the weld metal in all spatial positions of the weld pool in the range of significantly smaller mean values of the main technological parameters;
- to form a good conditions for transfer of every droplet of electrode metal into the molten pool. This makes it possible to reduce sputtering of electrode metal from 20% to 3% as a result of controlling the energy parameters of the welding;
- to increase the rate of weld pool solidification in 2 – 3 times as a result of the nonstationary energy effect of heating source on the weld pool with decreasing the temperature of molten metal;
- to decrease the degree of residual strains in the welded structures;
- to improve the quality of the welded joints and deposited surfaces (to improve the formation of the weld in all spatial positions, the structure of the weld metal and HAZ. This is determined by the controlled solidification of the weld pool. This is accompanied by the intensification of the hydrodynamic processes in the molten pool resulting in a more uniform distribution of the alloying elements through the entire volume of molten metal and intensive weld pool degassing;

- to improve mechanical properties of the welded structures: the size of the HAZ is reduced and the structure of the weld metal refined. This is of considerable importance for repeated loading.

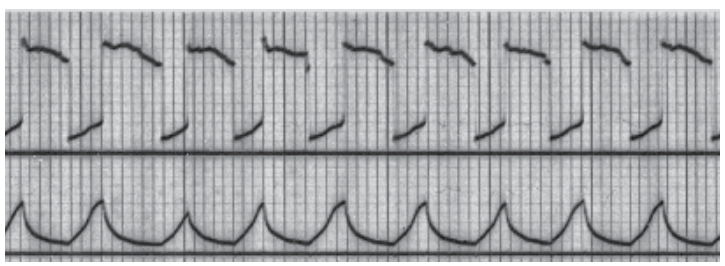
The important advantage of pulsed welding is the ability to stabilize the instantaneous values of main parameters in the stages of melting and transfer of an electrode metal droplet.

2. Quality of welded joint

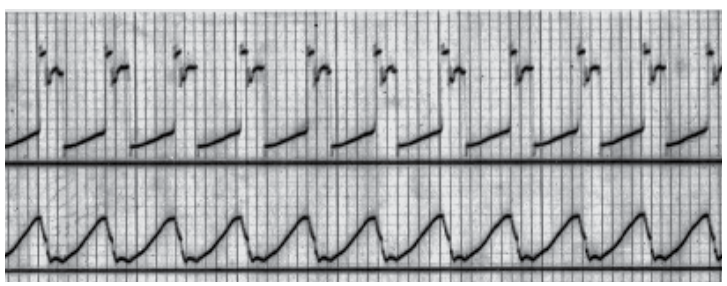
The main problem in welding in different spatial positions of high-quality inspected welded structures (joints in transmission pipe-lines, containers for oil and gas, chemical industry, boiler and power equipment, components of road-building machinery, equipment in the industry of engineering materials), operating under different types of loading at a subzero temperature, is to ensure the required quality of root, filling and capping layers and high mechanical properties of the welded joint. Up to 90% of defects, detected in the inspection of the quality of welded joints, are associated with defects in the root layers of welded joints, for example: undercutting, lack of fusion, nonmetallic inclusions or pores. The main reason for the formation of these defects, in addition to those associated with low quality of preparation, is the disruption of the welding conditions (welding speed, arc voltage, current), and that the regimes are not adhered to an optimum values.

Conventional welding processes can ensure the required quality of welded joints only in the case of efficient preparation of the welded joint and with the use of high-quality materials.

The above disadvantages can be eliminated by providing the welding process energy parameters constant in time, or varying them by a certain program.



a)



b)

Fig. 1. Oscillograms of current (lower curve) and voltage (upper curve) of: a) unstabilized and b) stabilized processes of CO_2 welding

The pulsed technologies are more efficient from the viewpoint of controlling the formation of welded joints in the presence of a large number of perturbing factors (defects in assembly of the joints, low quality of electrode materials, changes in the spatial position of the weld pool, variation of mains voltage, etc.). These are characterized by a stable penetrating capacity of the arc on the level of the instantaneous values of current and voltage with only slight dependence on the quality of electrodes.

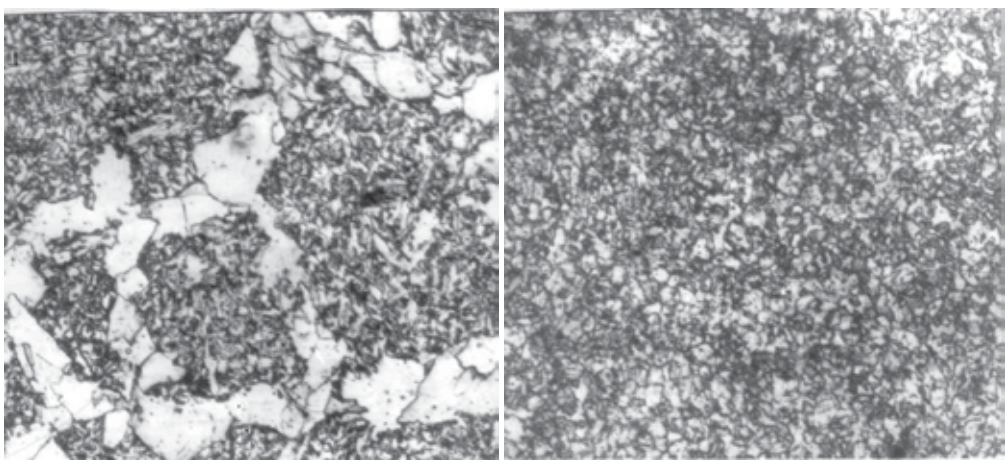
Fig. 1 shows oscillograms of the stabilised pulsed arc CO_2 welding process using Sv-08G2S wire and the conventional process without stabilisation of the energy parameters.

Primarily, this relates to one-sided pulsed-arc welding of root joints with the formation of the reversed bead without any backing strip and welding on reverse side in all spatial positions. The welding speed reaches 20 - 30 m/hr, whereas in uphill welding it is no more than 5 – 7 m/hr (Saraev & Shpigunova, 2002).

The technology of pulsed welding in different spatial positions is greatly simplified, the required properties and service reliability of welded joints are easily achieved, and the quality parameters of the welded joints improve: the size of the HAZ and zone of overheating near the surface of weld is reduced and the size of the normalized ferrite grain decreases (Fig. 2).

Transition to the pulsed regime of variation of the energy characteristics in surfacing makes it possible to control the processes of solidification in the weld pool and HAZ and decrease the degree of burnout of alloying elements from the weld pool. This is determined by the restriction of the time during which they are held at the high-temperature of the melt of the weld pool and by the increase of the rate of solidification of the weld pool. This is accompanied by the intensification of the hydrodynamic processes in the weld pool resulting in a more uniform distribution of the alloying elements through the entire volume of molten metal.

The application of adaptive pulsed welding of low-alloy steels results in formation of more dispersed and homogeneous structure of welded joint, than in welding by a permanently burning arc. The effect takes place in all layers of welded joints: root, facing, filling (Shpigunova & Glazunov, 2008 a).



a) b)

Fig. 2. Structure of the welded joint in: a) stationary and b) pulsed regimes of welding of 12Х1МФ steel, ×500

3. Optimal algorithms of pulsed control of the energy parameters of the welding process

The main purpose of computer aided design of pulsed technology is the development of an optimum algorithm of control over all links of technological chain from effective using of dynamic properties of power sources and programmed variation of the arc heat output to the HAZ microstructure changes, which provide required strength properties of welded joints and hard-facing coatings. The complexity of the problem is the necessity of welding phenomena studying from the viewpoint of kinetics of melting, thermodynamics, physical metallurgy of welding, the theory of heat conduction, hydrodynamics, the plasma theory, strength theory. Computer aided design of pulsed arc welding technology permits to solve such technical problems as the creation of new materials with preset thermo-mechanical and strength properties.

Extensive use abroad is made of welding with algorithms of pulsed control of the energy parameters of the process, as a rule, on the basis of a strictly defined programme. In this case, the main energy characteristics of the welding arc are calculated in advance and are set in strict accordance with the varied technological parameters (feed rate of electrode wire, open circuit voltage of the power source, etc.) These processes, such as: inert gas welding, non-consumable-electrode arc welding, plasma-arc welding can be used efficiently in the absence of perturbing influences on the object of automatic control.

The important advantage of pulsed welding is the ability to stabilize the instantaneous values of main parameters in the stages of melting and transfer of an electrode metal droplet. Such parameters as: the arc voltage; duration of typical stages of micro cycle - arcing time in the pulse t_{pulse}, the break with the duration t_p; instantaneous and mean values of current; arc power in a separate microcycle; melting energy of every electrode metal droplet can be the main controlled parameters of adaptive pulsed technological process.

Adaptive algorithms of pulsed control are corrected, during a technological process, through channels of feed backs in relation to the variation of the instantaneous values of the main energy characteristics of the welding process (arc current and voltage, peak short-circuiting current, arc power in a separate microcycle, melting energy of every electrode metal droplet). This makes it possible to supply more efficiently the energy required for melting and transfer of every droplet of electrode metal, control weld formation, taking into account its spatial position, reduce deformation of the welded joint by regulating the heat input in the welding and surfacing zone.

These processes take place with minimum deviations of the instantaneous values of the energy characteristics of the process, so that it is possible to calculate with sufficient accuracy the moment of separation and transfer of every electrode metal droplet to the molten pool and ensure detailed examination of the processes in the "power source - electrode - arc - molten pool" electrodynamic system as in a single object of automatic control.

The realization of the algorithms of pulsed control in current welding and surfacing equipment is associated with the introduction of additional sections and units into the structure of equipment. The units are introduced both into the circuits for controlling the output parameters of the power system and directly into the welding circuit (Fig. 3).

Selection of a specific technical solution depends on solving the technological problems and is determined by the frequency range of the algorithms of pulsed control of the energy parameters of the welding and surfacing processes.

- The following frequency ranges of the algorithms of pulsed control are selected:
- 5000 ÷ 100 Hz – for increasing the stability of arcing and decreasing the size of transferred droplets;
- 100 ÷ 25 Hz – for controlling the transfer of electrode metal in all spatial positions;
- 25 ÷ 0,25 Hz – for improving the formation of the welded joint in all spatial positions as a result of decreasing the size of the weld pool and increasing the rate of solidification;
- from 0,25 Hz and lower – for controlling the solidification processes in the weld metal and the HAZ (Fig. 2).

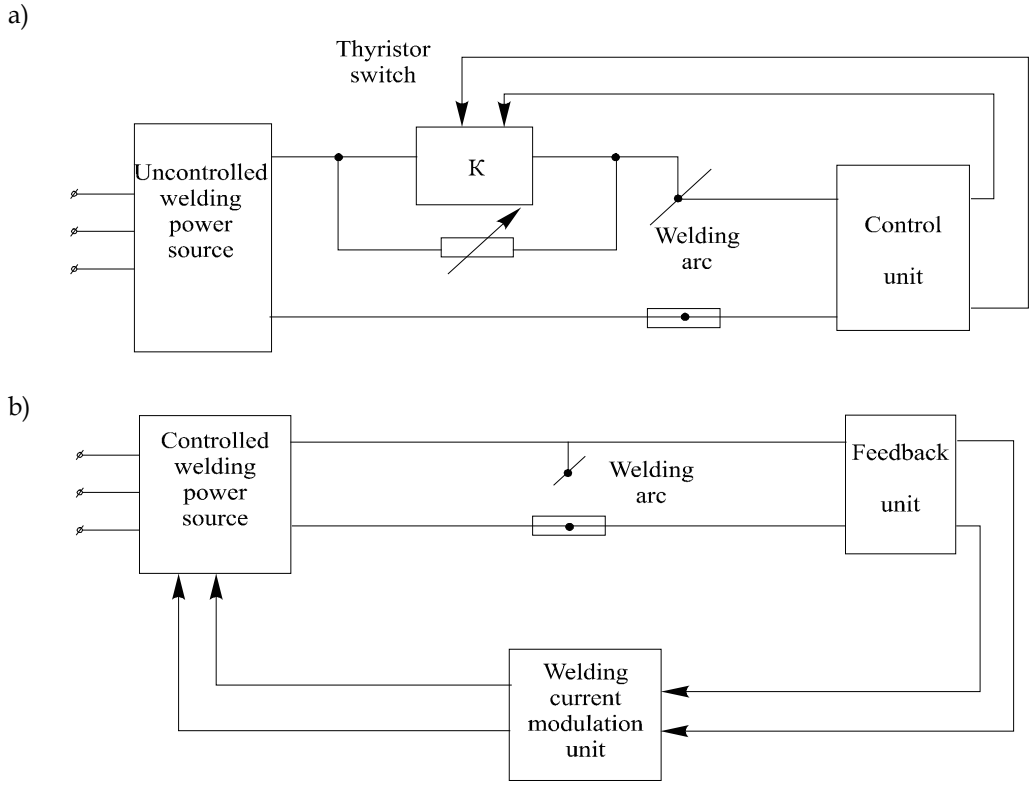


Fig. 3. Systems realizing adaptive pulsed technological processes of welding with:
a) uncontrolled and b) controlled power sources

The most complicated electrical engineering problem is the development of regulators for the frequency range 25 ÷ 5000 Hz. This is associated with the fact that sections of the regulator must ensure a very short restoration time of the controlled properties. In practice, this approach can be realized by introducing into the structure of the power supply system special high-current semiconductor regulators capable of switching large pulsed currents of 1000 A or even higher (Fig. 3 a).

Development of regulators in the frequency range 25 ÷ 0,25 Hz and lower is possible on the basis of semiconductor elements with low and medium power. As a result of the relatively long duration of current pulses, they can be shaped through the channels of phase control of

welding rectifiers and through the power circuit of the excitation windings of welding generators (Fig. 3 b) (Loos et al., 1998).

The developed technology of single-sided arc welding of root welds with formation of the reversed bead by the modulated current using coated electrodes is based on the special algorithm of control over the energy parameters, which permits to form during the technological process a condition of the welded zone as the result of pulsed arc action, when the components of melted electrode coating are intensively displaced beyond the forming permanent joint. Such an approach allows supply the formation of root welds without additional backing strips by electrodes of any coating, including the main type, to use coated electrodes manufactured in Russia instead of expensive imported electrodes. Well-known in a world practice the welding technological processes of root welds in condition of free formation (without additional backing strips) are based on application of special electrodes with a thin coating, that limits the fields of application of the given technologies. The proposed technology gives the possibility of downward welding of vertical welds that significantly increases the welding speed and simplifies welding technique in various spatial positions for a welder of lower qualification.

A large amount of experience has been accumulated in the last decade with the application of mechanized CO_2 welding in the production of metal structures in different spatial positions. The experience of production trials, however, has revealed a number of disadvantages related to defects in welded joints, lacks-of-fusion along the edges and between the layers due to instability of the welding process, and continuous change of the weld pool position in space. The above disadvantages can be eliminated by providing the welding process energy parameters constant in time, or varying them by a certain program. The optimal algorithms of control of the energy characteristics of the process, developed by computer-aided design methods, and specialized equipment (UDGI-201UKhLZ thyristor regulator) permit conducting the technological process of single-sided pulsed arc welding of the root welds with reverse bead formation without additional backing or backing run welding from the inside in CO_2. The using of UDGI-201UKhLZ regulator makes it possible to stabilize the welding process as result of fine-droplet transfer of electrode metal into the weld pool with the minimum 2 - 3% splashing of electrode metal in the range 70 - 200 A in mechanized and automatic welding with electrode wires with a diameter of 0,8 - 1,2 mm; simplifies welding technology in all spatial positions in the presence of large variations of the gap between the welded edges; increases 3 - 4 times the productivity of welding operations as a result of ensuring the possibility of downward welding. The speed of downward welding runs into 20 – 30 m/hr and upward welding speed is no more than 5 – 7 m/hr. The characteristic lack of penetration of downward welding, as a result of the weld pool inleakage in traditional CO_2 welding methods, is absolutely absent.

Fig. 4 shows typical oscillograms of such process. The proposed technological process has additional regulation parameters: t_i – arcing time in the pulse and t_{p11} – time of the break introduced at the moment of rupture of the liquid bridge. These parameters in accordance with the adaptation scheme are able to automatically correct the energy parameters of welding regime in relation to the perturbing influences so that it is possible to stabilize the heating and energy indicators of the process. The stability of such a welding process predetermines a stable quality of weld formation which also depends on the short-circuiting frequency $f_{s.c.}$, the holding time of the liquid droplet on the electrode tip, the droplets size and uniformity of their transfer.

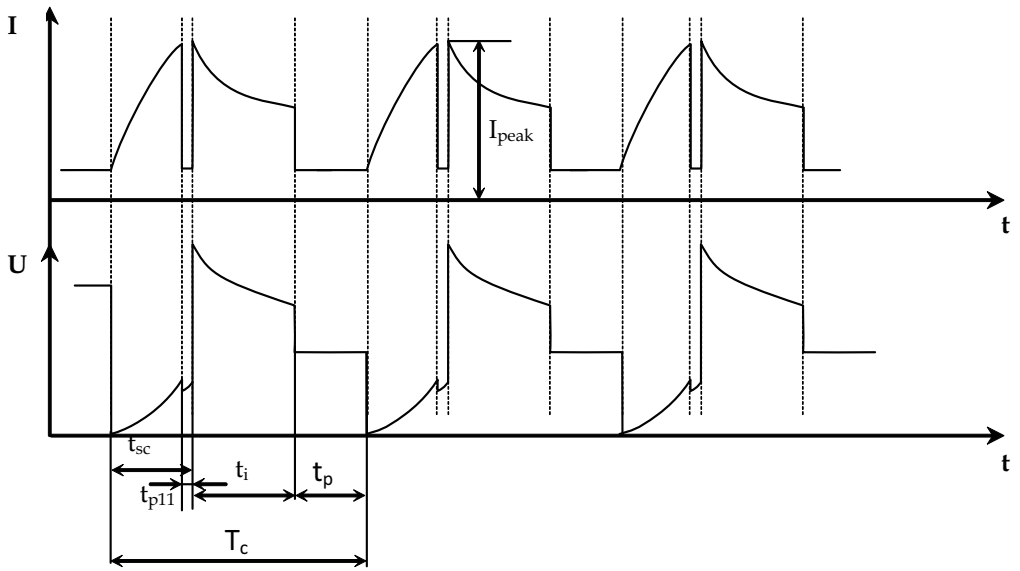


Fig. 4. Oscillograms of current I and voltage U of adaptive pulsed arc CO_2 welding

This set of process parameters can be optimized at the stage of technological preparation of production, in order to produce a sound welded joint operable under different types of loading in cold climate region. The results of research of the developed models of melting and metal transfer with systematic short-circuiting of the arc gap during the pulsed welding process, using a computer experiment, permits: to evaluate the influence of technological and energetic parameters complex of the process on the penetration of the weld metal, the shape and sizes of the weld and heat-affected zone; to predict strength properties and quality of welded joints (Shpigunova & Saraev 2003).

4. Mathematical modelling of heat and mass transfer in pulsed arc welding by melting electrode

4.1 Physical simulation of pulsed arc welding with forced short-circuiting of the arc gap

It is necessary to provide complex investigation of the welding arc physics and the electromagnetic processes in welding power source. The principle of metal transfer "one drop per pulse" is realized in adaptive pulsed arc welding in CO_2 medium.

The block-scheme of the power supply of adaptive pulsed arc welding is shown in Fig. 5.

The examination will be based on one of the control algorithms examined in (Saraev & Shpigunova, 1993).

The period of arcing in the pulse (Fig. 6) is characterized by rapid melting of the electrode tip under the welded component. As a result of the force effect of the arc, the weld pool metal is displaced into the tail part and is maintained there throughout the entire melting stage. After this period of arcing, the welding current in the pulse is increased in steps to the value of the background current. This results in a corresponding decrease in the melting rate of the electrode and a weakening of the force effect of the arc on the weld pool which tries at this moment to fill the crater formed below the electrode tip in the stage of the current pulse. Together with this effect, the electrode metal droplet tries to occupy a position, coaxial with

the electrode, mainly as a result of a decrease in force of reactive pressure of release of the gas, and also due to the forces of the weight of the droplet and surface tension.

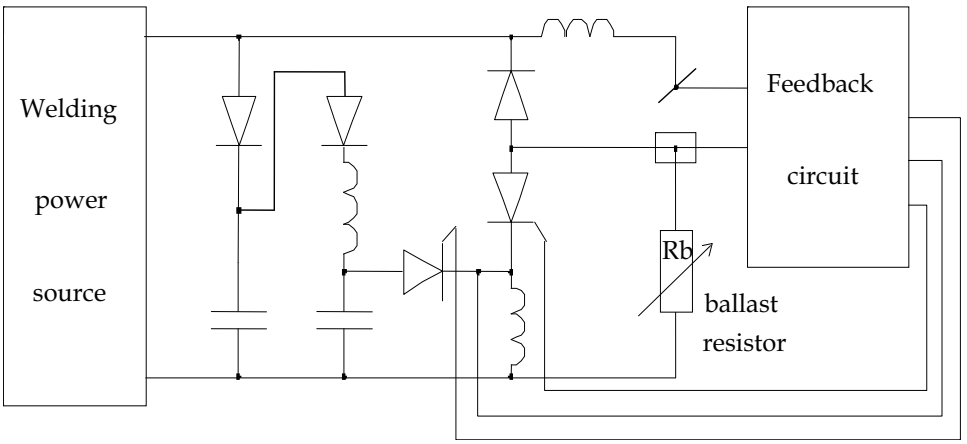


Fig. 5. Block-scheme of the power supply of adaptive pulsed-arc welding process

Forced short circuiting takes place as a result of these counter movements, and the initial moment of the short circuit is characterized by an increase in current in the welding circuit which increases along an exponent determined mainly by the interactive resistance of the smoothing choke coil. With this mechanism of electrode metal transfer, the formation of a stable bridge between the electrode and the weld pool is achieved in the first stage of short circuiting. This greatly increases the rate of increase of the short circuit current and, at the same time, accelerates the formation and fracture of the liquid bridge. In the short circuit stage, the transfer of electrode metal into the weld pool is accompanied by an increase in voltage (also in the case of the avalanche-like increase of current). This indicates the irreversibility of fracture of the bridge, as a result of a stepped decrease in current.

The entire period of short circuiting is characterized by the fact that the controlling effect in acceleration of failure of the bridge is played by the electrodynamic force which tries to "squash" the electrode metal along the melting line, separate the electrode metal droplet and apply to it the accelerating "pulsed force" for movement in the direction of the weld pool.

The final stage of fracture of the bridge (approximately 10^{-4} sec prior to the moment of arc reignition) is accompanied by the dominant effect of the surface tension force. However, as a result of the short duration of the given period, its contribution to the fracture of the liquid bridge is negligible. The duration of the break is set either parametrically, or in relation to the condition of the arc gap in the given stage.

After completion of the break, increasing current, the electrode starts melting in the pulse current period. Subsequently, the course of the process is identical with that described previously.

Such mechanism of controlled transfer of electrode metal into the weld pool is operating in the realization of other adaptive algorithms of the pulsed control of the energy parameters of the process. The only difference is that the perturbation effects, determined by the droplet transfer of electrode metal and the special features of formation of the weld metal in different spatial positions, operate in different stages of the welding microcycle, depending

on the variation of the arc gap length at the start of the effect of the pulse current or the integrated value of high-voltage in the stage of parametrically specified background period up to the moment of fracture of the bridge, or when the force effect of the arc on the weld pool in the period of the current pulse is determined in relation to the duration of the break prior to a short circuit, indicating the ability of the weld pool during changes of its special positions (Saraev, 1994).

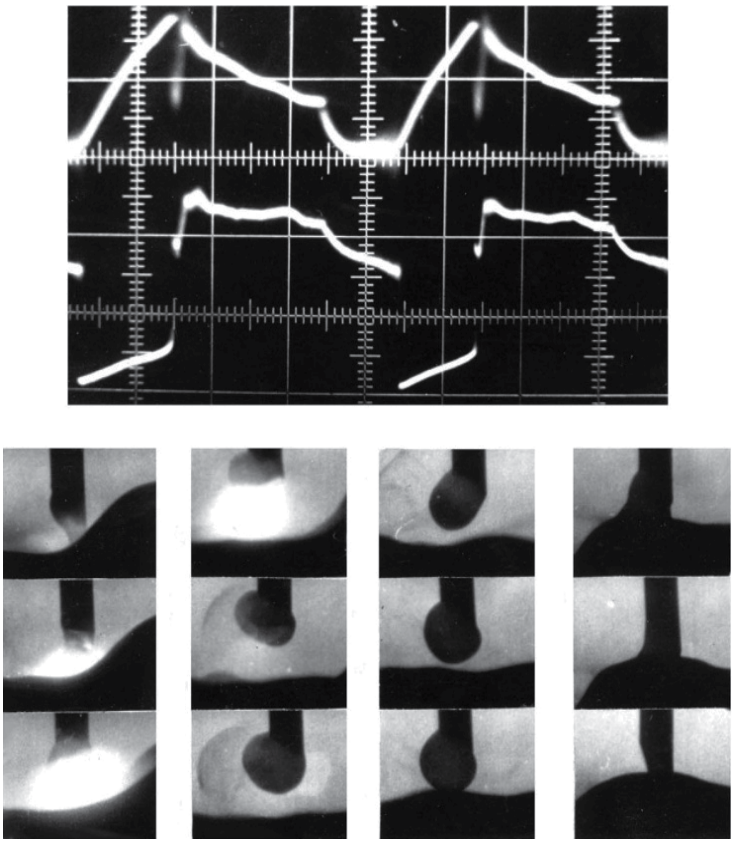


Fig. 6. Oscillograms of current (upper curve) and voltage (lower curve) and film frames of microcycle of CO_2 welding with forced short-circuiting of the arc gap

The results of analyzing the oscillograms and experimental data obtained by high-speed filming of pulsed-arc welding process in CO_2 with SV08G2S wire ∅ 1,2 mm (Fig. 6) make it possible to specify the following main features of the pulsed process and formulate a number of assumptions for mathematical modeling of such a process:

- the molten pool moves with specific periodicity in such a manner that prior to every short-circuit, the molten pool occupies the same position in relation to the continuously fed electrode. Therefore, in calculations, these movements can be ignored;
- the break current prior to a short-circuit is low and has no marked effect on melting of the electrode in the break period;
- the break introduced at the moment of arc reignition does not affect the thermal processes in the system and, consequently, its effect can be ignored in the calculations;

- electrode metal formed at the electrode tip as a result of melting of the continuously fed electrode has the form of a spherical segment;
- thermophysical constants (α,$c\gamma m$), used in calculations, do not depend on temperature, where α is the temperature coefficient of resistance; $c\gamma$ is the volume heat capacity of electrode wire; m is the latent heat of electrode melting which takes into account transition from one aggregate state to another;
- the resistance of electrode extension R_e depends on both the temperature to which it is preheated T_p and the steel grade.

Every microcycle T_c consists of the three typical stages (Fig. 6, Fig. 7):

1. short-circuiting with the duration $t_{s.c.}$;
2. arcing in a pulse with the duration t_{pulse};
3. the break prior to a short-circuit, duration t_{pause} (I_{peak} is the peak value of the short circuit current).

The simplified mechanism of droplet formation and electrode metal transfer to the molten pool can be described as follows.

After rupture of a bridge, the energy build-up in the choke coil during a short-circuit generates in the arc gap and rapidly melts the electrode. At the initial moment, the melting rate of the electrode V_e is higher than the feed rate V. Consequently, the width of the arc gap increases. Part of the molten electrode metal, which remains at the end from a previous microcycle, rapidly increases in the volume at the start to a hemisphere with the diameter $2R_e$ and then to a spherical segment with the height h. When welding current is reduced and the volume of the spherical segment increase the burn-off rate decreases and the width of the arc gap slightly decreases. After completion of the arcing process in the pulse and a reduction of welding current to the break current I_o, the burn-off rate rapidly decreases and the arc gap closes up as a result of continuous electrode feed. A short-circuit takes place, during which metal is transferred to the molten pool.

In accordance with the described mechanism of growth of the electrode metal droplet, the volume of the spherical segment in the second period increases at the rate dh/dt in the direction of the continuously fed electrode with the speed $\boldsymbol{V}$. This is accompanied by countermovement of the melting line of the electrode with the melting speed $\boldsymbol{V_m}$.

4.2 Mathematical modelling of heat and mass transfer in welding with systematic short-circuiting of the arc gap

Taking into account these special features of the pulsed process and the assumptions, a cyclogram of welding current I and voltage U, as well as a simplified diagram of growth of the droplet of molten electrode metal and the shape of the finite weld are shown in Fig. 7 and Fig. 8, respectively.

The object of our research is a mathematical model of melting and transfer of electrode metal with systematic short-circuiting of the arc gap in carbon dioxide medium on the base of algorithm of control, shown in Fig. 7 (Saraev & Shpigunova, 1993).

There are a large number of investigations (Dyurgerov, 1974), (Popkov, 1980), (Lebedev, 1978) which have been carried out to describe mathematically the power source – welding arc system in welding with systematic short-circuiting of the arc gap using the mean parameters of the conditions. However, they did not reflect the technological stability of the process, because a deviation of one of these parameters within the limits of a separate microcycle leads to its disruption. In particular, when welding in different spatial positions, the deviation resulting in an increase of a specific parameter, such as the peak short-circuit

current within the limits of the microcycle, leads to splashing of the metal from the molten pool during arc reignition. The variation of electrode stick-out results in a change of the heat generated in the stick-out, which in turn affects the energy balance of the arc, etc. In this case, the amount of heat generated in the extension may reach 15,6% of the entire arc heat (for low-carbon electrodes), which is equal to 55% of the heat required for melting the electrode. In rapid heating of the electrode with passing current, the burn-off rate of the electrode increases.

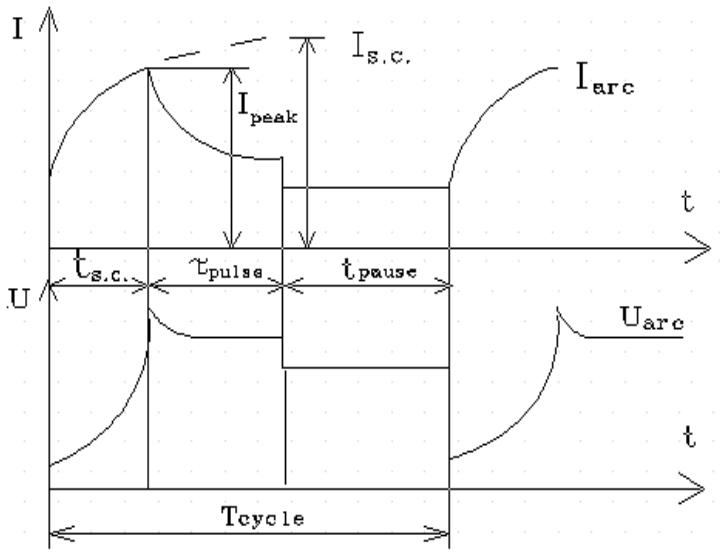


Fig. 7. The cyclogram of current I and voltage U for the power source – welding arc system

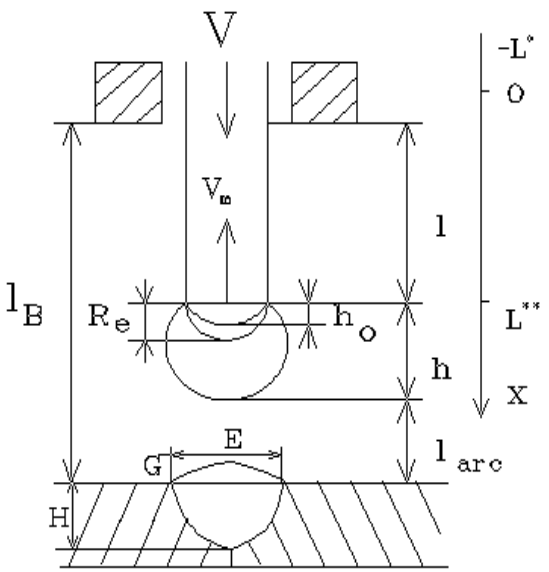


Fig. 8. The scheme of growth of a droplet of molten electrode metal and shape of the finite weld

The temperature distribution along the electrode is defined on the base of solution of heat conduction equation in consideration of convective heat exchange to space. Therefore, an examination is made of one-dimensional heat propagation in cylindrical electrode bar, fixed in current lead tip, within the limits of statement of a problem from (Saraev & Shpigunova, 1993). The interval on which a function is defined changes from $-L^*$ (L^* = const) to $L^{**}(t)$ in axis OX (Fig. 8).

$L^{**}(t)$ - is the length of the unmelted part of the heated electrode extension in moment of time t,

L^* - is the part of electrode with temperature gradient from T (in the point x = 0) to T^* (in the point x = $-L^*$) and convective heat transfer coefficient $\tilde{\alpha}$. In interval from x = 0 to x = L^{**} the arc and passing through the electrode current are a heat sources. There are no internal heat sources in interval from x = 0 to x = $-L^*$.

It is necessary to note the following: the electrode is moving with the speed $\boldsymbol{V}$ (Fig. 8) that means position change concerning to the current lead tip and upper boundary. This is equal to the regular feed of the "cold" mass. The lower boundary is moving with speed $\boldsymbol{V_g = V - V_m}$, where $\boldsymbol{V_m}$ - the melting speed of lower end of electrode as a result of the heat release from the arc.

Heat conduction equation is solving within the limits of statement of a problem (Saraev & Shpigunova, 1993). It means that the amount of heat flow on the lower boundary of the electrode and value of passing through the electrode current are determined by the problem solving from paper (Saraev & Shpigunova, 1993) in every time moment. The electrode resistance in interval from x = 0 to x = L^{**} and melting speed are determined subject to the temperature distribution calculated from the heat problem solution.

4.3 Heat conduction equation with boundary conditions:

Thus:

$$\frac{\partial}{\partial t}(\gamma \cdot T) = \frac{\partial}{\partial x}\left(\frac{\lambda}{c} \cdot \frac{\partial T}{\partial x}\right) + \frac{\rho \cdot I^2}{c \cdot F^2} - \frac{\alpha \cdot P}{c \cdot F} \cdot \left(T - T^*\right) - \frac{\partial}{\partial x}(V \cdot \gamma \cdot T) \tag{1}$$

Note, that:

ρ - specific resistance,

$$\rho(T) = \rho * (1 + \alpha * \Delta T) \tag{2}$$

$$\Delta T = T - T * \tag{3}$$

ρ^* - specific resistance at T^*,

α^*- temperature coefficient of resistance.

Here:

t - time,

γ - the density of electrode material,

T - temperature,

T_m - melting temperature,

T_d - drop temperature,

λ - thermal conductivity,

I - current,

c - specific heat of electrode material,

α- convective heat exchange coefficient,

F - cross-section area of electrode,
P - electrode perimeter.
The boundary conditions for the short circuit interval and arcing in a pause interval are:

$$T(-L^*,t) = T^* \tag{4}$$

$$T\left(L^{**}(t),t\right) = T_m \tag{5}$$

For the arcing in a pulse interval:

$$T(-L^*,t) = T^* \tag{6}$$

$$\frac{\partial}{\partial x} T\left(L^{**}(t),t\right) = -\vec{q} \tag{7}$$

The amount of heat flow - $\boldsymbol{q}$ on the lower (moving) boundary of solution field get from the law of conservation of heat energy (*Mathematical Modelling*, 1979):

$$Q^+ = Q_1^- + Q_2^- + Q_3^- \tag{8}$$

Where: Q^+ - heat quantity from arc;
Q_1^- - heat quantity consumable to electrode melting;
Q_2^- - heat quantity consumable to the increase in molten metal temperature from $T = T_m$ to $T = T_d$;
Q_3^- - heat quantity transferred for a depth into metal.
The complete version of the Eq. 8 is:

$$U_a^e \cdot I \cdot F^{-1} = \left[M + C \cdot \left(T_d - T_m\right)\right] \cdot \gamma \cdot V_m - \lambda \frac{\partial T}{\partial x} \tag{9}$$

Where:

$$V_m = dL_m / dt\,;$$

U_a^e - effective anode voltage,
M - specific heat of melting.
Another condition on moving boundary is that its temperature approximately equal to the temperature of melting:

$$T\left(L^{**}(t),t\right) = T_m \tag{10}$$

Let's develop moving differential grid. Melting speed is determined by iterations.
Discrete analogue of Eq. 1 is developed according to digitization method (Patankar, 1984) and “check volumes” method and solved by the run method.
Therefore, there is the system of differential equations for each interval of microcycle:
"short circuit" (Fig. 7, Fig. 8):

$$\frac{\partial I}{\partial t} = \frac{U_{xx} - R_s I(t)}{L_G} \tag{11}$$

$$R_s = R_\Sigma + R_{L,} \qquad R_L = F^{-1} \cdot \sum_{i=1}^{N-1} \rho \cdot \Delta x_i$$
$$dL = V \cdot dt \tag{12}$$

$$t_{s.c.} = \frac{L_G}{R} ln\left(\frac{U_{xx} - R_s \cdot I_0}{U_{xx} - R_s \cdot I_p}\right) \tag{13}$$

$$I_p = 1,5 \cdot 10^5 \cdot h + 154 \tag{14}$$

T(x,t) is determined from Eq. 1 for all intervals.
For "arcing in a pulse":

$$dL = (V - V_m)\, dt, \quad l_g = l_b - (L + h)$$
$$\frac{dh(t)}{dt} = 2r^2 \frac{V_m}{h^2 + r^2} \tag{15}$$

$$\frac{dI}{dt} = \frac{U_{xx} - (U_{ak} + \beta \cdot l_g) - R_s \cdot I}{L_G} \tag{16}$$

For "pause":
$I = I_0$,
h = const,
$t_p = l_g/V$,

$$dL = V \times t \tag{17}$$

Here:
U_{ak} - anode-cathode voltage,
β - gradient of voltage of arc column,
L_G - inductance of welding circuit,
U_{xx} – open circuit voltage of the power source,
R_s - resistance of welding circuit,
r - radius of electrode.

4.4 Results of computer simulation

The system of non-linear differential equations for each interval of microcycle is realized by means of numerical methods in a computer. To solve a system of non-linear differential equations, the authors used an explicit two-step method of the predictor – corrector type of the second order of accuracy on smooth functions. Since the model process must be cyclic (output parameters of a single microcycle represent input parameters for the next microcycle), the problem was solved by an iteration approach. The criterion for convergence of the process is the difference ΔI of the values of the current curve on adjacent iterations: ΔI ≤ 0.01%. Original software for realization of such problems have been developed (Shpigunova et al., 2000; Shpigunova & Glazunov, 2008 b).
The results of numerical solution of the problem give the full information about object of control at each time moment: the value of current $I(t)$, arc voltage $U(t)$; the size of the drops transferred from the electrode $h(t)$; the preheat temperature of the electrode extension $T(L,t)$;

the length of the arc gap l_g; the resistance of the unmelted part of the heated electrode extension R_L, and so on; permit to determine the interrelation between energetic characteristics of the pulsed arc welding process ($I(t)$, $U(t)$), sizes of weld and HAZ (Fig.) with the most important regulated technological parameters of the process (V - electrode feed rate, L - electrode extension, U_{xx} - open circuit voltage of the power source, t_{pulse} - arcing time in the pulse, t_p – time of pause, frequency of transferred droplets of electrode metal) and to give the quantitative assessment.

Fig. 9 shows temperature distribution in electrode with length L (mm) at different time moment of microcycle for following values of the thermophysical quantities and parameters of the process of CO_2 pulsed welding with Sv-08G2S wire: L = 12 mm, t_{pulse} = 10 ms, $t_{s.c}$ = 2.16 ms, U_{xx} = 45 V, V = 0.222 m/sec, β = 3.6 V, L_G = 0.00018 H, I_0 = 20 A, r = 0.5 mm, T_d = 2673 K, U_{ak} = 22 V, $c\times\gamma$ = 5.23×10^6 J/m^3×K, λ = 39.65 W/m×K, α* = 0.003 1/K, α= 100 W/m^2×K, R_s = 0.05 Ω.

Every temperature curve consists of two ranges: range of smooth change of temperature as a result of heat release by passing current and range of quick increasing of temperature as a result of heat input by arcing. The depth of heat penetration by arc depends on the speed of melting front motion very much (Fig. 9).

Fig. 10 shows melting speed of electrode depending on time moment of microcycle for different values of U_{xx} – open circuit voltage of the power source and V = 0.138 m/sec.

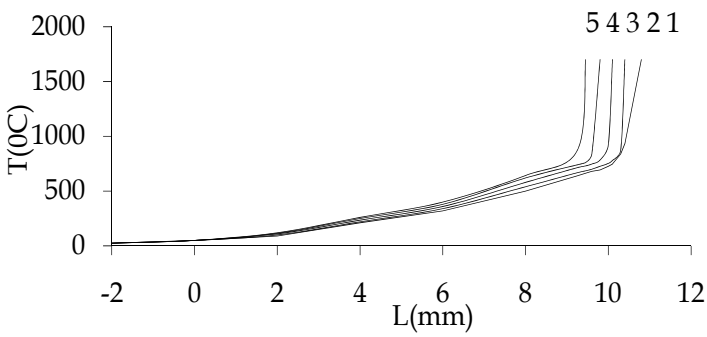


Fig. 9. The temperature distribution in electrode extension at different time moment: 1 – t_1 = 2.16 msec; 2 – t_2 = 4.66 msec; 3 – t_3 = 7.16 msec; 4 – t_4 = 9.66 msec; 5 – t_5 = 12.16 msec

The examined pulsed technological process is characterised by the fact that its parameters can be regulated over a wider range than the stationary process. This is possible because, in addition to the generally accepted regulation parameters of the welding process (open circuit voltage of the power source $U_{o.c.}$, electrode feed rate V, electrode extension l_b), there is an another parameter: arcing time in the pulse which, combined with the general parameters, makes it possible to control the dimensions of the transfer droplets and their frequency. In addition, the regulating capacity of the power source – welding arc system is controlling the welding process and compensating different perturbing influences on the regulation object, i.e. the arc.

For example, when the electrode extension is varied in the range 8 ÷ 20 mm, the temperature to which the electrode extension is heated rapidly increases. This may be compensated by a corresponding increase of the arcing time in the pulse. It is thus possible to stabilize the mean value of welding current and, consequently, the electrode burnoff rate.

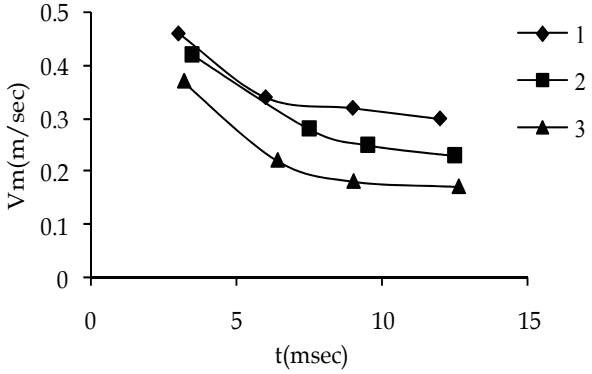


Fig. 10. Dependence of melting speed on time of microcycle at different values of open circuit voltage of the power source: 1 – U_{xx} = 45 V; 2 – U_{xx} = 40 V; 3 – U_{xx} = 35 V

Important technological parameters of the welding process are the frequency of transferred droplets of electrode metal and their volume, which determine to a large extent the required geometrical dimensions of the weld. These parameters can also be mutually compensated in accordance with the required ranges.

For example, an increase of the electrode extension reduces the frequency of short-circuits and increases the volume of molten electrode metal within the limits of a separate microcycle. These parameters can be maintained in the required ranges by reducing the arcing time in the pulse. This increases the frequency of short-circuiting and reduces the volume of molten metal (Fig. 11, Fig. 12).

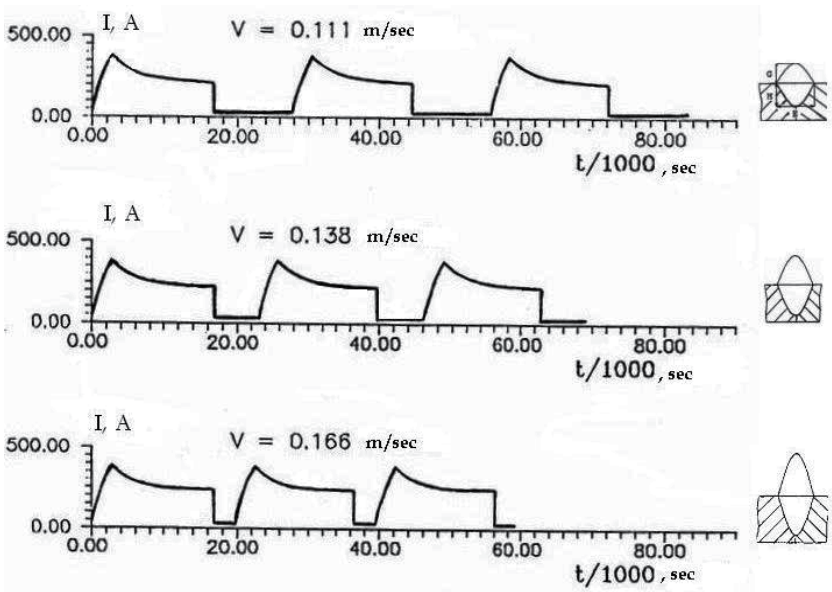


Fig. 11. Calculated cyclograms of welding current for different values of electrode feed rate V and correlative weld shape (H, E, G) for every complex of technological parameters of pulsed arc welding. $U_{o.c.}$ = 41 V, t_{pulse} = 7.5 ms, L = 12 mm, V = 0.110 ÷ 0.220 m/sec, I_0 = 30 A

This approach makes it possible to calculate the cyclograms of welding current as a result of computer experiments for wide range of values of regulated technological parameters for the CO_2 pulsed welding with Sv-08G2S wire: d = 0.8 ÷ 1.2 mm, L = 8 ÷ 12 mm, t_{pulse} = 5 ÷ 18 ms, $t_{s.c}$ = 2.16 ms, U_{xx} = 35 ÷ 45 V, V = 0.111 ÷ 0.222 m/sec (Fig. 11).
An increase of the electrode feed rate increases the frequency of short-circuits at almost constant instantaneous values of current in both in the short-circuit range and the arcing time range in the pulse. This results in a higher stability of the welding process, as well as constant dimensions of the transferred droplets of electrode metal irrespective of the spatial position of the molten pool. This is of considerable importance for maintaining stable parameters of the welding process.

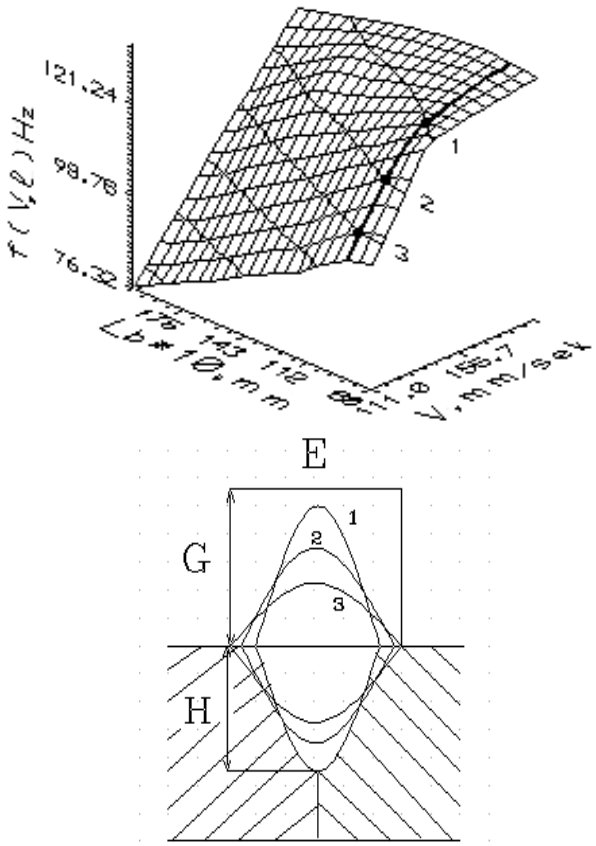


Fig. 12. Dependence of frequency of short-circuits $f_{s.c.}$ on electrode extension L_b at different values of electrode feed rate V ($U_{o.c.}$ = 35 V, t_{pulse} = 5 ms) and correlative weld sizes (H, E, G) for different complex of regulated technological parameters of pulsed arc welding ($f_{s.c.}$, L_b, V)

Fig. 12 shows the dependence of frequency of short-circuits $f_{s.c.}$ on electrode extension L_b at different values of the electrode feed rate V for open circuit voltage of the power source $U_{o.c.}$ = 35 V, pulse time t_{pulse} = 5 ms and the dependence of weld shape (H – penetration depth, E – weld width, G – throat) on complex of technological parameters of CO_2-shielded pulsed-arc welding (weld shape 1, 2, 3 correlate to complex of technological parameters 1, 2, 3).

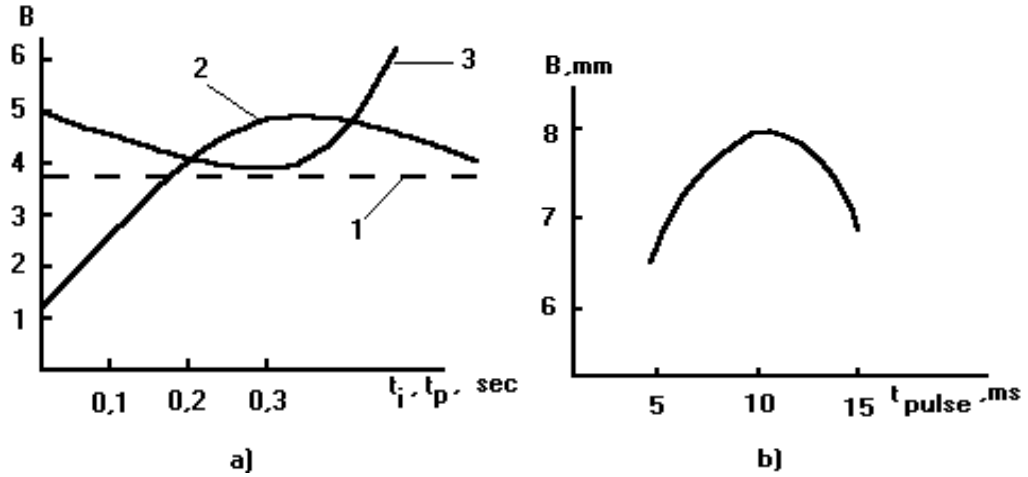


Fig. 13. Dependence of HAZ width on time of pulse (t_i) and time of pause (t_p) in pulsed arc welding:

a) by modulated current - experimental results for I_{mean} = 120 A:

1 - stationary arc;

2 – t_i – var, t_p = 0.3 sec;

3 – t_p – var, t_i = 0.3 sec.

b) pulsed arc welding in CO_2 of low carbonaceous steel – computer experiment (721° C isotherm)

4.5 Control of weld formation

During designing of optimum algorithm of control over pulsed regime of welding there is need to choose such combination of welding parameters (automatic welding, semiautomatic, submerged arc welding, and welding in an atmosphere of shielding gases): U_a, I(t), j - current density in electrode, v - welding speed, chemical compositions (the marks), granulation of flux, type of the current, its polarities which provide formation of the joints with proper sizes, shape and quality with high operating strength. The sizes - depth of penetration H (Fig. 12), breadth of weld E, height of deposited bead G, and shape of weld are determined by quantity of the heat transferred to the article and by the character of its introduction. Under the effect of high-speed heat source the penetration area (the area restricted by the isotherm of melting T_m):

$$F_p = \frac{1}{E \cdot c \cdot \gamma \cdot T_m} Q_p$$

$$Q_p = \frac{0.24 \cdot I(t) \cdot U_a \cdot \eta}{v}$$

Here:

$Q_p = Q/v$,

Q - arc power,

η - effective efficiency of arc.

Using of typical coefficients: $\psi_p = E/H$ coefficient of penetration form, $\psi_f = E/G$ coefficient of strengthening form and constants: A, K , μ - obtaining from experiments for low-alloyed

steel and low-carbon steel during the welding in CO_2 medium by electrode wire Sv-08G2S and Sv-08GS gives the possibility to apply the equations connecting typical sizes of weld and energetic characteristics:

$$H = A \cdot \sqrt{Q_p / \psi_p}$$

$$\psi_p = k(19 - 0{,}01I)\frac{d \cdot U_a}{I}, \qquad E = \psi_p \cdot H$$

Where d – is the diameter of electrode.
During hard-facing or welding of butt joints without edge level with zero clearance the deposited metal is in the form of the bead above the sheet's surface, therefore $G = F_n / \mu * E$, where μ - is the coefficient of bead completeness, F_n - the area of cross-section of deposited bead.
Using of this dependencies for research of the effect of the main technological and additional regulated parameters: $U_{o.c.}$, I_p, l_b, V, t_i, $f_{s.c.}$, t_p, *h* on the sizes of the given welded joint during welding and hard-facing on the base of computer experiment for wide range of values of technological and energy parameters of welding regime (Fig. 12) it is possible to design optimum regime, which provide required relationships of geometric sizes of the weld ψ_p, ψ_f for the given type of the welded joint, which characterize its technological and operating strength.
So, in automatic and semiautomatic welding with $\psi_p < 0.8$ the joints inclined to the hot crack formation are produced, with $\psi_p > 4$ - too wide welds with small penetration depth, what is inefficient from the viewpoint of arc power using and the result is deformation increasing. For the well formed welds the optimum range of values is $\psi_f = 7 - 10$. The narrow and high welds with small ψ_f do not have smooth connection with basic metal and have dissatisfied ability to work under variable loads. The large values ψ_f correspond to wide and low strengthening, what is undesirable because of decreasing of weld section in comparison with basic metal section because of the vibrations of molten pool level.

5. Conclusion

The results of analyzing the cyclograms of welding current and oscillograms show that they qualitatively coincide. The deviation of the calculated instantaneous values of welding current from the experimental data does not exceed 10%. This convergence level makes it possible to recommend the proposed mathematical model and original software for it numerical realization for examining actual pulsed technological processes.
The proposed mathematical model of melting and transfer of electrode metal in welding with systematic short-circuits of the arc gap and original software for it realization takes into account the heat generation in the electrode extension (heat conduction and convective heat exchange to space).
The action of heat processes in electrode on speed of electrode melting and amount of transferred molten metal, the nature of formation and transfer of every electrode metal droplet, and the state of the arc gap on the level of instantaneous values in limits of mathematical model (Saraev & Shpigunova, 1993) have been investigated.
There is most difference in temperature distribution, melting speed and sizes of transferred electrode metal droplet from paper (Saraev & Shpigunova, 1993) near the melting front or in time moment $t = t^0 + t_{s.c.}$.

The area of a solution existence for the proposed model is bigger than for model (Saraev & Shpigunova, 1993).
The developed approach permits on the basis of numerical realization of developed models to solve the inverse problem - to design optimum algorithms of control over the system "power source – arc – weld pool" in pulsed welding process, to determine optimum complex of values of regulated parameters depending on solving of technological problem, such as: decreasing of molten metal sputtering, improvement dynamic properties of power sources, the formation of weld with preset sizes and service properties.
This set of pulsed process parameters can be optimizing at the stage of technological preparation of production, in order to produce a sound welded joint operable under different types of loading.
The results of researching of the developed mathematical models of melting and metal transfer with systematic short-circuiting of the arc gap during the pulsed welding process, using a computer experiments, permits: to evaluate the influence of technological and energetic parameters complex of the process on the penetration of the weld metal, the shape and sizes of the weld and heat-affected zone.
Using these mathematical models permit to reduce the volume of experiments, aimed at developing pulsed conditions and to predict the strength properties, quality, reliability and operating longevity of welded joints.
Physics-mechanical and chemical processes of the formation of primary crystalline structure of weld and HAZ are multiple and difficult to simulation. There are a large number of accompanying factors which in particular cases may be a cause for control over welded joint strength in welding technology. The thermo-capillary convection applies to this class of phenomena. It leads to effect of irregular distribution of impurity concentration in melt that entails a change of crystallization front and affects the formation of structure of welded joint. Also it is necessary to examine the diffusive mechanism of impurity migration and the kinetics of polymorphous transformation.
The developed methodology of computer aided design of advanced technologies, which suppose the creation of integral model of adaptive pulsed process of welding and hard-facing; modeling; original software; adaptive algorithms of pulsed control and special equipment are most effectively used for defectless welding of root joints with the formation of the reversed bead in all spatial positions without any additional backing strip and welding on the reverse side by electrodes of different types.
The use of specialized equipment for developed pulsed methods of welding makes it possible to stabilize the welding processes as a result of fine-droplet transfer of electrode metal into the weld pool with the minimum (2 - 3%) splashing of electrode metal in the range 70-200 A in mechanized and automatic welding with electrode wires with a diameter of 0.8 – 1.2 mm; ensure guaranteed high-strength properties of important welded joints to be subjected to 100% inspection; simplifies welding technology in all spatial positions in the presence of large variations of the gap between the welded edges; increases 3 – 4 times the productivity of welding operations as a result of ensuring the possibility of downhill welding.
The regions of application of advanced pulsed welding technologies are: the welding of root, filling and facing layers of ship structures in different spatial positions and butt joints in the processing and transmission pipelines with a diameter of 32 – 1420 mm; boiler and power equipment for important applications, welding robotic technological systems for engineering companies.

6. References

Saraev, Y. & Shpigunova, O. (2002). Adaptive pulsed arc methods of welding of high-responsible welded structures for the service under cold climate extreme conditions, *Proceedings of 1st Eurasian Symposium EURASTRENCOLD-2002,* Vol. 2, pp. 42 - 49, Yakutsk, July 2002, Publ. IPTPN SB RAS, Yakutsk

Shpigunova, O. & Glazunov, A. (2008 a). Ensuring the quality of technological process of welding on the basis of a registration of energy parameters for producing defectless welded joints. *Izvestiya Vysshikh Uchebnykh Zavedenii. Fizika,* № 8/2, (2008), pp. 304 - 306, ISSN 0021-3411

Loos, A.; Lukutin, A. & Saraev, Y. (1998). *Power Sources for Pulsed Electrical Engineering Processes,* Publ. Tomsk Polytechnical University, Tomsk

Shpigunova, O. & Saraev, Y. (2003). Computer aided design of pulsed arc welding ensuring defectless joint of metals. *Materials Science Forum,* Vols. 426 - 432 (August, 2003), pp. 4027 – 4032, ISSN 0255-5476, Trans Tech Publications, ISBN 0-87849-919-9, Switzerland

Saraev, Y. & Shpigunova, O. (1993). A mathematical model of melting and transfer of electrode metal with systematic short-circuiting of the arc gap. *Welding International,* Vol. 7, № 10, (1993), pp. 793 - 797, ISSN 0950 7116

Saraev, Y. (1994). *Pulsed Technological Processes of Welding and Surfacing,* Publ. Nauka, ISBN 5-02-030653-3, Novosibirsk, Russia

Dyurgerov, N. (1974). Reasons for periodic short-circuits of the arc gap when welding with a short arc. *Svarochnoe Proizvodstvo,* № 9, (1974), pp. 1 - 3, ISSN 0491-6441

Popkov, A. (1980). Stability of the power source – arc system when welding with systematic short-circuits of the arc gap. *Svarochnoe Proizvodstvo,* № 3, (1980), pp. 11 - 13, ISSN 0491-6441

Lebedev, A. (1978). Effect of heat generation in the electrode extension on the process of self-regulation of the arc. *Avtomaticheskaya Svarka,* № 7, (1978), pp. 9 - 11

Andrews, J. & McLown, R. (Eds.). (1979). M*athematical Modelling,* Publ. Mir, Moscow

Patankar, S. (1984). *Numerical Methods of Heat Exchange and Fluid Dynamics Problem Solving,* Publ. Energoatomizdat, Moscow

Shpigunova, O., Shpigunov, S. & Saraev, Y. (2000). Mathematical simulation of heat and mass transfer in pulsed arc welding by melting electrode. *Acta Metallurgica Sinica (English Letters),* Vol. 13, № 1, (February 2000), pp. 56 – 62, ISSN 1006-7191

Shpigunova, O. & Glazunov, A. (2008 b). Numerical simulation of pulsed arc welding by melting electrode. *Materials Science Forum,* Vols. 575 - 578 (April, 2008), pp. 786–791, ISSN 1662-9752, Trans Tech Publications, ISBN / ISBN-13 : 0-87849-392-1 / 978-0-87849-392-0, Switzerland

16

Mathematical Modelling of Structure Formation of Discrete Materials

Lyudmila Ryabicheva and Dmytro Usatyuk
Volodymyr Dahl East Ukrainian National University
Ukraine

1. Introduction

The mathematical modelling of different processes and events may be reduced, in most cases, to formulation of boundary-value problems for defined systems of differential equations. Series of statements and approximate methods for solving of such equations were developed by many authors. The most development have obtained variation methods, direct methods of mathematical physics and integral equation methods. These methods have specific capabilities and peculiarities, expanded class of observed problems, but were not completely eliminated most of principal contradictions. Nowadays, the most challenging method is finite element method (FEM). It has reached so high stage of development and popularity that can be no doubts of existence another approach competitive in capabilities and simplicity of realization (Segal et al., 1981; Wagoner & Chenot, 2001).
The advantages of finite element method are free selection of nodal points, arbitrary shape of region and boundary conditions, simplicity of generalization for different models of bodies and problems of any dimensionality, natural accounting the non-uniformity of properties and other local effects, using of standard programs for a whole class of problems. A finite element method is well grounded, the equivalence of its different forms to differential and variation formulations and, also, to special cases of Ritz method, Bubnov-Galerkin method and least-squares method established (Zienkiewicz & Taylor, 2000).
The first step of numerical solution is discretization of medium that allows reducing the problems with infinite number of degrees of freedom typical to continuous approach, to problems with finite number of unknown variables. Usually, discretization is including selection of certain number of nodal points with following implementation of two types of variables – nodal variables and special functions that are approximating the distributions of target parameters inside elements. In such case, the independent parameters are the nodal variables and distributions of target parameters that are determined by them (Zienkiewicz & Taylor, 2000).
During finite element approximation the integration procedure is replaced by more simple algebraic operators expressed through nodal variables by summation on elements. Partial differential equations are replaced by system of algebraic equations written for sequence of nodes and special functions by functions for finite number of nodal variables. The subsequent calculation of target values and determination of parameters of state may be executed by standard methods of numerical analysis. The general requirements for selection of finite elements and approximating functions are determined by convergence criterions of FEM (Zienkiewicz & Taylor, 2000).

The implementations of FEM to solving of technological tasks of plasticity theory and modelling of physical and mechanical properties associated with metal forming processes are described below. Large deformations specific for such processes are leading to changing the geometry of region and properties of material. In these cases most of peculiarities of plastic state that produce difficulties of numerical solution are appeared (Petrosjan, 1988; Wagoner & Chenot, 2001).

2. Solving of the non-stationary nonlinear coupled thermal-structural problem by finite element method

The behaviour of powder porous bodies at plastic deformation and high temperatures is characterizing by substantial non-uniformity that makes necessary application of numerical methods (Petrosjan, 1988). Nonlinear character of deformation and substantial non-uniformity of deformed state in combination with large temperature gradients are leading to the necessity of solving a non-stationary nonlinear coupled thermal-structural problem. The matter of this problem is that forming process of detail depends not only from degree of deformation and strain rate but, also, from temperatures which continuously changing by nonlinear laws (Wagoner & Chenot, 2001; Hallquist, 2006; Ryabicheva & Usatyuk, 2006). The sequence of solving of non-stationary nonlinear coupled thermal-structural problem consists of the followings steps: problem formulation, discretization scheme, computational procedure and computer visualization of results.

The eight node linear tetrahedron-shaped element has used for analysis of stress-strain state, temperature distributions and physico-mechanical properties. The fundamental idea is that five nodal points of element have common coordinates and each projection of their displacement is described by one equation (Hallquist, 2006). According to (Segal et al., 1981), a minimum of functional is corresponding to actual velocity field:

$$J = \iiint_V \sigma_{ij} e_{ij} dV - \iint_{S_k} p_i v_i dS, \tag{2.1}$$

where σ_{ij}, e_{ij} - are stress tensor and strain rate tensor;

p_i - are pressures applied on external border;

v_i - are velocities of displacements of points under the action of external forces;

V - is volume of body;

S_k - is surface of body.

During a finite-element approximation integration is replaced by summing up on elements and minimization of function (4.1) results in the system of equations:

$$[K]\{\dot{X}\} = \{p\}, \tag{2.2}$$

where $[K] = [K(X, \dot{X})]$ - is global stiffness matrix of system;

$\{p\}$ - are column-matrices of nodal velocities and forces.

Dependences between nodal velocities and strain rates and, also, stresses into element are looking like (Segal et al., 1981):

$$\{e\}^{(e)} = [B]\{v\}^{(e)}, \quad \{\sigma\}^e = [K]\{v\}^e. \tag{2.3}$$

Matrices [B] and [K] are determined by standard technique. The dependence between stresses and strain rates, determined by matrix [D], obtained using the following relation (Skorokhod, 1973; Segal et al., 1981; Shtern et al., 1982):

$$\sigma_{ij} = \beta\left[\phi e_{ij} + \left(\psi - \frac{1}{3}\phi\right)e\delta_{ij}\right], \tag{2.4}$$

where $\beta = \dfrac{\sqrt{1-\theta}\tau_0}{\sqrt{\phi\gamma^2 + \psi e^2}}$;

θ - is porosity of material;

ϕ, ψ - are porosity functions (Shtern et al., 1982): $\phi = (1-\theta)^2$, $\psi = \dfrac{2}{3}\dfrac{(1-\theta)^3}{\theta}$;

τ_0 - is ultimate intensity of deviatoric stresses for basic material of porous body.

The visco-plastic medium investigated during plastic deformation at high temperatures according to recommendations of Kachanov L.M. (Kachanov, 1969). A substantial metal flow is typical for visco-plastic medium at the certain load and flow velocity depends on viscosity of medium. In case of axis-symmetrical problem (Zienkiewicz & Taylor, 2000; Wagoner & Chenot, 2001):

$$[D] = \begin{bmatrix} \frac{4}{3}\phi + \psi & \psi - \frac{2}{3}\phi & \psi - \frac{2}{3}\phi & 0 \\ \psi - \frac{2}{3}\phi & \frac{4}{3}\phi + \psi & \psi - \frac{2}{3}\phi & 0 \\ \psi - \frac{2}{3}\phi & \psi - \frac{2}{3}\phi & \frac{4}{3}\phi + \psi & 0 \\ 0 & 0 & 0 & 2\phi \end{bmatrix}. \tag{2.5}$$

The kinetic equation of porosity changing in visco-plastic area looks like (Shtern et al., 1982):

$$\frac{d\theta}{dt} = (1-\theta)\left(\frac{\phi H\sigma}{\psi T} + \frac{\sigma}{\psi}\right), \tag{2.6}$$

where H - is intensity of shear strain rate;

σ - is current normal stress;

$T = \left(\dfrac{1}{2}\sigma'_{ij}\sigma'_{ij}\right)^{1/2}$ - is shear stress intensity.

Beginning of plastic flow corresponds to implementation of condition (Shtern et al., 1982):

$$f \equiv \psi T^2 + \phi\sigma^2 - \sigma_s^2 = 0, \tag{2.7}$$

where σ_s - is yield stress at linear tension $\left(\sigma_s = \sqrt{3}\Gamma_s\right)$.

System of equations (2.2) is algebraically nonlinear relatively to $\{\dot{X}\}$ and in relation to {X} it is a system of differential equations. The step-by-step loading method has used for its integration. In such case the displacement of deforming element is divided on the row of steps with value Δh . A nonlinear algebraic equations system (2.2) is solving during each of steps for determination of $\{\dot{X}\}$. The values equal to product of time step to average velocity between respective load steps are added to coordinates on previous time step for determination of nodal coordinates.

In case of time step size is quite small, the velocity distribution allows to define coordinates and deformation of nodal points at the end of step (Zienkiewicz & Taylor, 2000):

$$\begin{aligned} x_i(t+\Delta t) &= x_i(t) + V_i(t)\Delta t, \\ \varepsilon_{ij}(t+\Delta t) &= \varepsilon_{ij}(t) + \varepsilon_{ij}(t)\Delta t. \end{aligned} \tag{2.8}$$

Changes of shape and properties of material are calculating in such way and attained accuracy is usually sufficient for practical purposes.

The important feature of plastic deformation dependences is that they are not dependent directly on time. Therefore, a displacement of deforming element may be an internal time of system. An iteration process proceeds to stopping of change $\{\dot{X}\}$ and {X} with given accuracy. Changing of temperatures on the section of sample at high temperature deformation has determined using the heat conductivity law for each element.

Thus on each load step the analysis of interaction of contact surfaces for elements inside a sample or in contact with surface of instrument was executed, because contact interaction allows determination of heat conductivity only for inner and contacting elements (Segal et al., 1981; Wagoner & Chenot, 2001; Hallquist, 2006; Ryabicheva & Usatyuk, 2006).

The Fourier differential equation was implemented for heat conductivity analysis (Wagoner & Chenot, 2001; Ryabicheva & Usatyuk, 2006):

$$k_T\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2}\right)dV = C\rho\frac{\partial T}{\partial \tau}dV\,, \tag{2.9}$$

where k_T – is total coefficient of heat conductivity;

C – is specific heat capacity;

ρ – is density of material;

T – is temperature, K;

τ – is time of load step.

A minimum of heat conductivity functional is corresponding to each loading step (Wagoner & Chenot, 2001; Ryabicheva & Usatyuk, 2006):

$$Q = \iiint_V k_T\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2}\right)dV\,, \tag{2.10}$$

where k_T – is total coefficient of heat conductivity.

Integration of functional (2.10) is replaced by summing up on elements. A time step should be selected small enough in order to ensure homogeneous distribution of temperature and stationary heat transfer inside all elements. Solving of algebraic equation systems has

performed by Gauss method. A maximal change of parameters in any of elements should not exceed some value stipulated by strength properties of material.
Thus, the minimization procedure of functionals (2.1) and (2.10) for non-stationary, nonlinear and non-isothermal processes of deformation of powder porous body consists of solving of linear algebraic equation systems with verification of convergence criterion. The indicated procedure is repeating on each time step for all sequential stages of calculation.
The LS-DYNA 971 solver has used for solving the above mentioned problems.

3. Mathematical modeling and forecasting of mechanical properties of single- and multi-component powder materials

3.1 Mathematical model

The mathematical model of material that proposed for modelling of physico-mechanical properties of porous body is presented by system of constitutive equations that are describing physical and mechanical properties of components.
The finite elements that describe different components of materials are placed in a common mesh. It allows the possibility of taking into account interactions between components. The input data are volume fractions of components, their property in compact state, and also specified value of porosity. The elasto-plastic model of material is applied to all components. The independent parameters are nodal displacements (Segal et al., 1981).
The strain intensities ε_i and strain rates $\dot{\varepsilon}_i$ inside each element are defined through projections of nodal displacements onto the coordinate axes (Segal et al., 1981):

$$\varepsilon_{ix} = \frac{\sum_{\lambda=1}^{N} \frac{\partial u_x^\lambda}{\partial x}}{N}, \qquad \varepsilon_{iy} = \frac{\sum_{\lambda=1}^{N} \frac{\partial u_y^\lambda}{\partial y}}{N}, \qquad \varepsilon_{iz} = \frac{\sum_{\lambda=1}^{N} \frac{\partial u_z^\lambda}{\partial z}}{N},$$

$$\varepsilon_i = \frac{\sqrt{2}}{3}\sqrt{(\varepsilon_{ix}-\varepsilon_{iy})^2 + (\varepsilon_{iy}-\varepsilon_{iz})^2 + (\varepsilon_{iz}-\varepsilon_{ix})^2}, \quad \dot{\varepsilon}_i = \frac{d\varepsilon_i}{dt}. \tag{2.11}$$

where λ – is the node number;
N – is the number of nodes in a finite element;
$u_x^\lambda, u_y^\lambda, u_z^\lambda$ – are projections of nodal displacements onto the coordinate axes;
$\varepsilon_{ix}, \varepsilon_{iy}, \varepsilon_{iz}$ – relative deformations of finite element onto the coordinate axes.

Taking into account the thermo-mechanical coefficients, the Cowper and Symonds equation for stress intensity σ_i inside a finite element looks like (Hallquist, 2006):

$$\sigma_i = \left[1 + \left(\frac{\dot{\varepsilon}_i}{C}\right)^{\frac{1}{p}}\right](\sigma_0 + \beta E \varepsilon_i). \tag{2.12}$$

where σ_0 – is the initial yield stress of a component;
E – is the Young modulus;
$\beta = k_t k_\upsilon k_\varepsilon$ – is the hardening coefficient of component;
C, p – are arbitrary constants.

From the condition of equality of resultant displacements follows that after meshing of finite elements with different properties to common mesh, values of stress intensity, deformation intensity and strain rate at neighbour elements describing different components of material will be different. It means that values of σ, ε, E, Poisson's ratio v and density ρ in the given area of sample may be expressed in the following way (Ryabicheva & Usatyuk, 2007):

$$\sigma = \frac{\sum_{j=1}^{n} \sigma_j}{n}, \quad \varepsilon = \frac{\sum_{j=1}^{n} \varepsilon_j}{n}, \quad E = \frac{\sigma}{\varepsilon}, \quad v = \frac{\varepsilon_{xy}}{\varepsilon_z}, \quad \rho = \frac{\sum_{j=1}^{n} \rho_j}{n \sum_{i=1}^{m} \delta_i}. \tag{2.13}$$

where n – is the number of finite elements in a given area;

ε_{xy}, ε_z – are the radial and axial deformations;

m – is the number of components in the material;

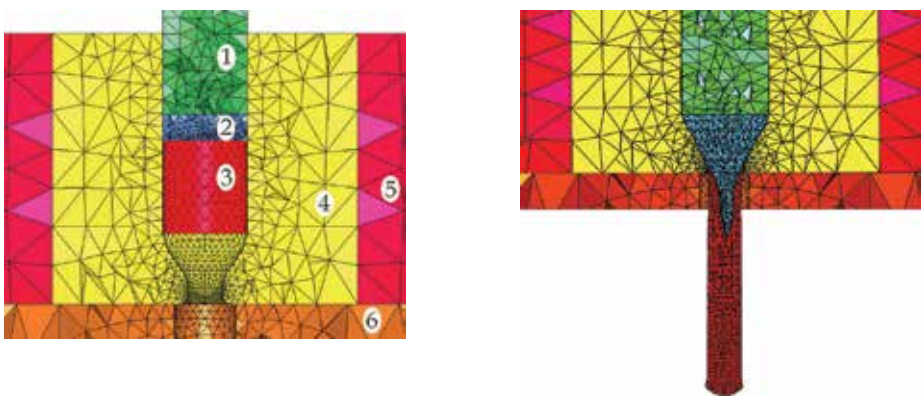
δ_i – is the volume fraction of component.

It is significant that in the proposed model porosity is described as a component of powder material and zero-elements are used for its modelling. The volume fraction of zero-elements is equal to given porosity of the material.

3.2 Initial data

The distributions of stress intensity, degree of deformation, strain rate, temperature and density at the deforming process, estimation of quality of manufactured items have been performed during mathematical modelling of extrusion of rod-shaped billet with predetermined complex of mechanical properties.

The porous fibrous sample with density 8.75 g/cm^3 obtained by pressing of copper fibres with diameter 0.8-1.3 mm and 6-12 mm length have used as initial billet. The finite element model of extrusion of porous fibrous pressing is presented on Fig. 3.1, a. A cylindrical graphite press-washer 2 for filling out the cavity of working part of matrix 4 at the end of extrusion was placed between punch 1 and initial billet 3 for removing finished product from a matrix without butt-end (Fig. 3.1, b).



a) b)

Fig. 3.1. The finite element model of extrusion: a - is the initial position; b – is the operation-terminating position: 1 - is the upper puncheon; 2 - is the press-washer; 3 - is the initial pressing; 4 - is the matrix; 5- is the bandage; 6- is the lower plate

The temperature on the beginning of extrusion is 920^{0}C, friction coefficient is 0.15. The diameter of porous fibrous pressing is 23.7 mm, height - 30 mm. The density of graphite press-washer is 2.2 g/cm^3. The diameter of calibrating hole in the matrix was equal to 12.9 mm, 9.1 mm and 6 mm, the reduction ratio was 3.6, 7.3 and 16.8, respectively. A detailed analysis of stress-strain state was performed in three sections passing through the beginning (Fig. 3.2, section 1-1), middle part of deformation zone (Fig. 3.2, section 2-2) and output of matrix 4 (Fig. 3.2, section 3-3).

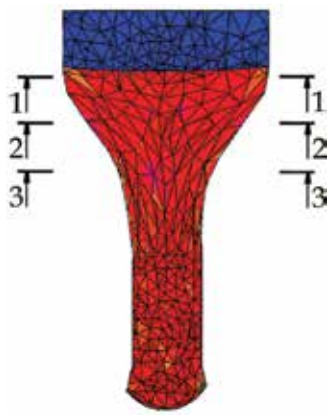


Fig. 3.2. The investigated sections

3.3 Modelling of stress-strain state and distribution of temperatures during extrusion

The stress-strain state picture is almost the same with all reduction ratio investigated, however, at $\lambda = 16.8$ the values of stress intensity and hydrostatic pressure are much higher then at $\lambda = 3.6$ and 7.3 (Fig. 3.3, a, b). In such conditions the distribution of stress intensity by section of pressing from axis to wall of matrix is more uniform. Its maximal value 145 MPa was reached at the output of deformation zone near the wall of matrix. The existence of gradients of additional stresses, tensile stresses near the walls of matrix and compression stresses in the inner layers of metal leads to complex character of hydrostatic pressure changing by section of billet. The value of hydrostatic pressure has grown up and become 1380 MPa (Fig. 3.3, b).

Obviously, the maximal point at radius of billet r = 2 - 4 mm is corresponding to beginning formation of flow-through flaw in the billet, that is well concordant with one of basic laws of metal forming theory about the flow of metal in the direction of least resistance – by the axis of matrix and, also, corresponding to distribution of strain intensity (Fig. 3.3, d).

The presence of tensile deformations in central part of sample ensures larger value of strain intensity that diminishing to the walls of matrix due to the influence of friction. Increasing of longitudinal tensile normal stresses from axis to wall of matrix causes decreasing of transversal layers thickness near the wall and their thickening at the central area of billet. The strain rate intensity in sections 1-1 and 2-2 has conditioned by proximity of certain volumes to elastic zones of cylindrical segment of container and calibrating segment of matrix. It should be noted that difference between strain rates in sections 2-2 and 3-3 becoming lower with growing of reduction ratio that testifies increasing of stiffness of stress-strain state while increasing of reduction ratio.

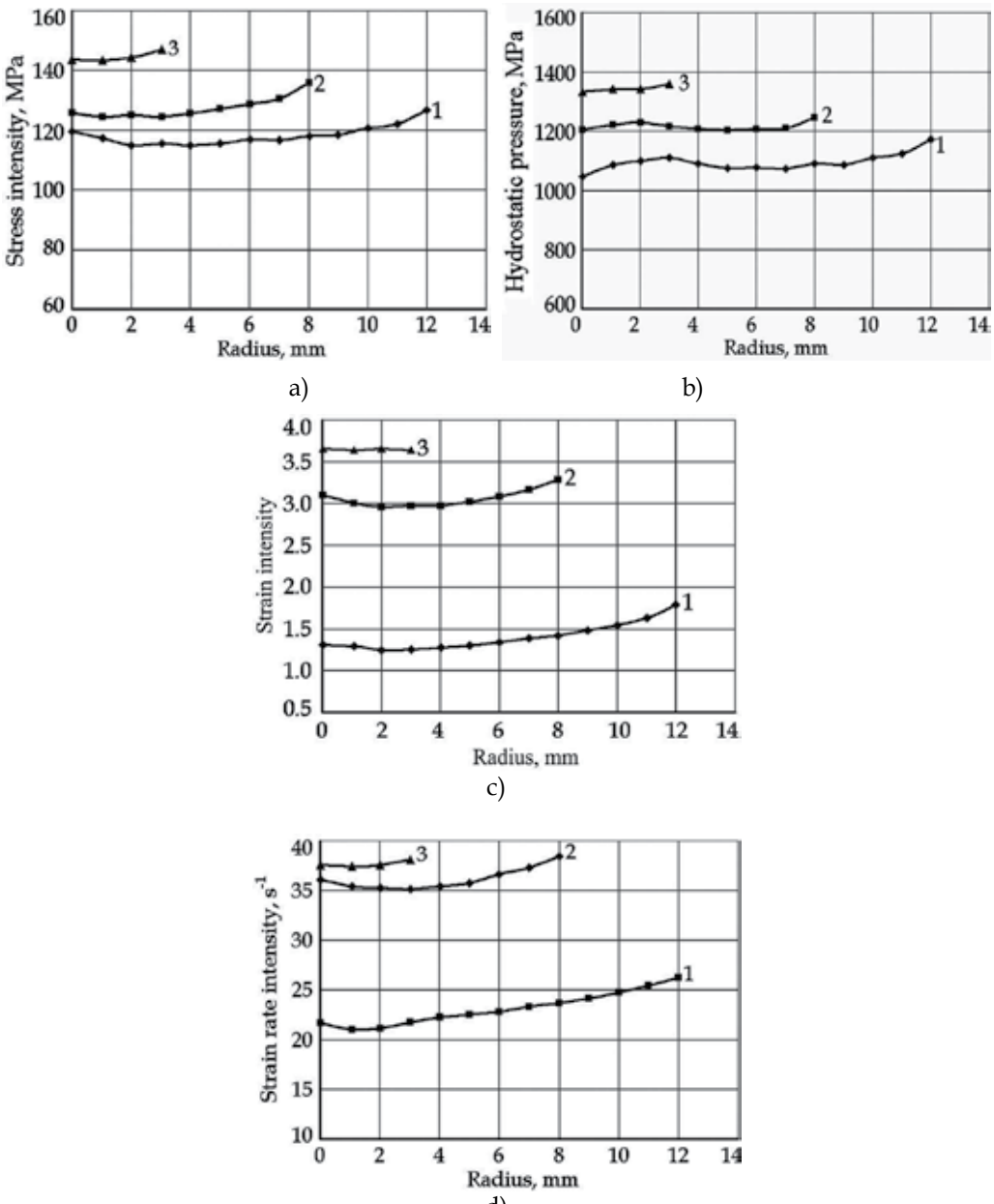
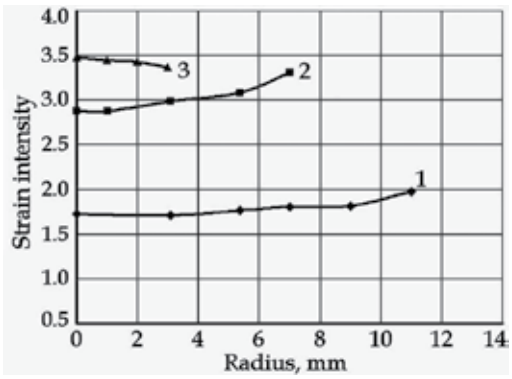
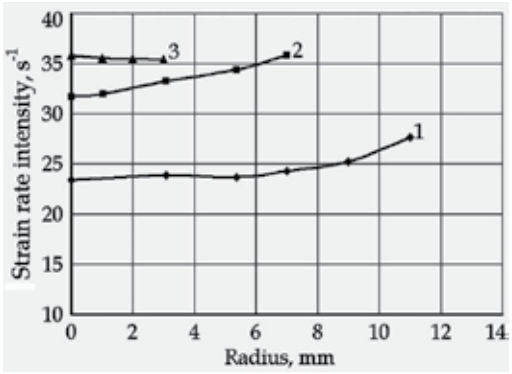


Fig. 3.3. The distribution of stress intensity (a), hydrostatic pressure (b), strain intensity (c), strain rate intensity (d) at λ=16.8: 1– is the section 1–1; 2– is the section 2–2; 3 – is the section 3–3

Computer modelling of stress-strain state during extrusion of fibrous pressing is corresponding to results of analysis of common scheme of changing the coordinate grid by its state in the beginning, middle and the end of deformation zone in experimental investigation (Fig. 3.4).



a)



b)

Fig. 3.4. The distribution of strain intensity (a) and strain rate intensity (b) by sections:
1 – is the section 1–1; 2 – is the section 2-2; 3 – is the section 3-3

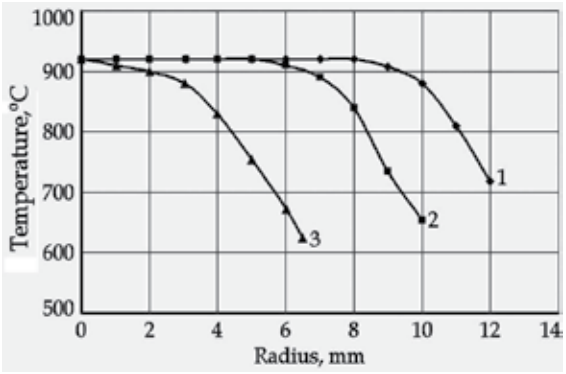
The maximum values and most uniform distribution of strain intensity and strain rate intensity have reached at section 3-3 that ensures production of sample of given diameter. The distributions of temperatures for all of three reduction ratios into investigated sections are similar (Fig. 3.4).

It should be noted that temperature goes down in section 1-1 only in the 3 mm layer of pressing due to heat transfer to the matrix at all of three reduction ratios. However, the temperature decreases more intensively to 650 °C at $\lambda = 16.8$ (Fig. 3.5). Decreasing of temperature in the centre of deformation zone (section 2-2) goes more intensively due to growth of reduction ratio that is related to increasing of contact area of pressing with walls of matrix. The most rapidly it appears in section 3-3 when at the small diameter of article happens sharp falling of temperature by whole section. The reasons of such temperature changes are heat conductivity processes in layers of pressing at extrusion and between pressing and walls of matrix.

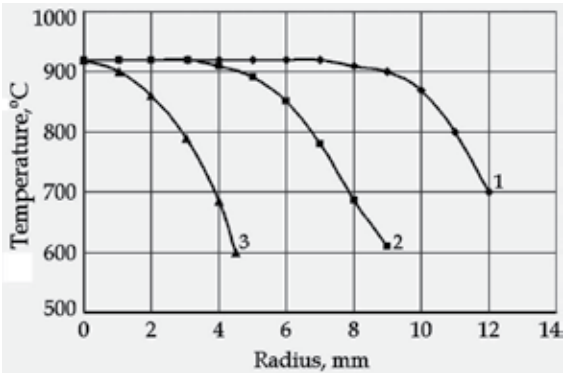
The distribution of density at different reduction ratios is presented on Fig. 3.6. The density is falling down while increasing the distance from centre of sample to circumference of the sample. Specifically, at $\lambda = 3.6$ the density fell to 8.70 g/cm^3, at $\lambda = 7.3$ to 8.87 g/cm^3. The

density is slightly decreasing to 8.93 g/cm^3 at $\lambda = 16.8$ and almost constant by section of sample.

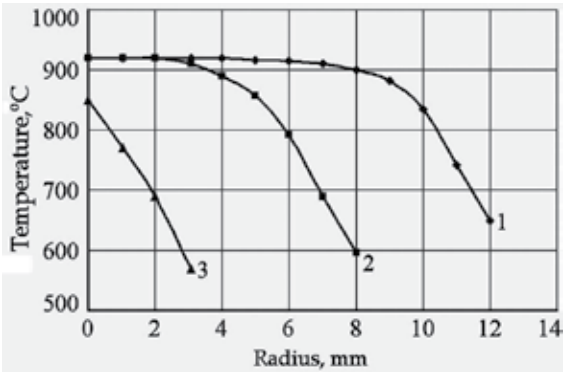
The shear stress intensity, that is growing up while increasing of reduction ratio, have defined for estimating the consolidation of fibres at current density of samples (Fig. 3.7).



a)



b)



c)

Fig. 3.5. The distributions of temperatures by sections of billet during extrusion: $\lambda = 3.6$ (a), $\lambda = 7.3$ (b), $\lambda = 16.8$ (c): 1 – is the section 1–1; 2 – is the section 2–2; 3 – is the section 3–3

A comparison of shear stress intensity performed with critical shear stress τ_{cr} determined by formula:

$$\tau_{cr} = \frac{\sigma_T}{\sqrt{3}}, \tag{3.1}$$

where σ_T - is the yield stress at given temperature and strain rate conditions.

It should be noted that at $\lambda = 3.6$ the intensity of shear stress is lower than critical shear stress, at $\lambda = 7.3$ the value of τ is a bit lower than τ_{cr} that testifies to incomplete consolidation of fibres, and at $\lambda = 16.8$ its value much higher than τ_{cr}. A high hydrostatic pressure within 1000-1380 MPa at the reduction ratio $\lambda = 16.8$ ensures full consolidation of fibres at extrusion and production of nonporous fully consolidated material that meeting the requirements of standard. These data have been verified by mechanical properties of material obtained experimentally.

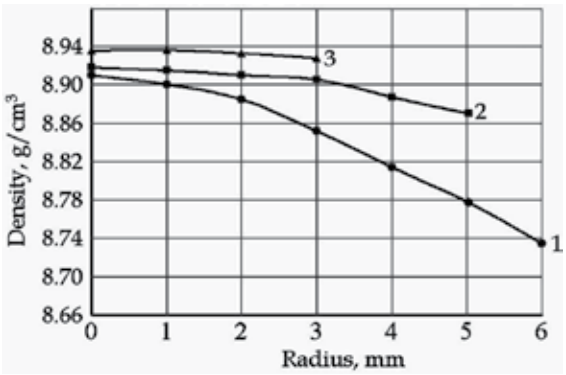


Fig. 3.6. Distributions of density by sections of copper sample: 1 - λ=3.6; 2 - λ=7.3; 3 - λ=16.8

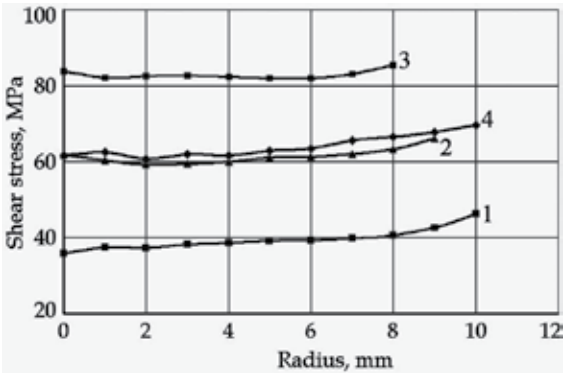


Fig. 3.7. Shear stress intensity: 1 – λ = 3.6; 2 – λ = 7.3; 3 – λ = 16.8; 4 – is a critical shear stress

Thus, modelling of direct extrusion of initial fibrous pressing with the density of 8.75 g/cm^3 has shown that density conformed to density of compact material obtained at the reduction ratio 16.8 ensuring complete consolidation of fibres. However, finite element simulation

allowed identifying defects of material flow similar to experimental results (Fig. 3.8). It has established that flow-through flaw appears on upper end of sample at all reduction ratios.

a) b)

Fig. 3.8. The flow-through flaw on after end (a) and loosening on exposed face (b) of samples obtained from fibrous pressing

Evolution of flow-through flaw at the reduction ratio 16.8 is presented on Fig. 3.9. The flow-through flaw does not appear during the initial stages of deformation (Fig. 3.9, a, b) while metal did not fill in the working segment of matrix. A flow-through flaw nucleates at transferring of metal to deformation zone into the centre of pressing (Fig. 3.9, c). A slight increasing of hydrostatic pressure on its edges observed. A flow-through flaw spreads deep into billet by the end of extrusion (fig. of a 3.9, d, e) and its depth l_{sk} is depending on the reduction ratio.

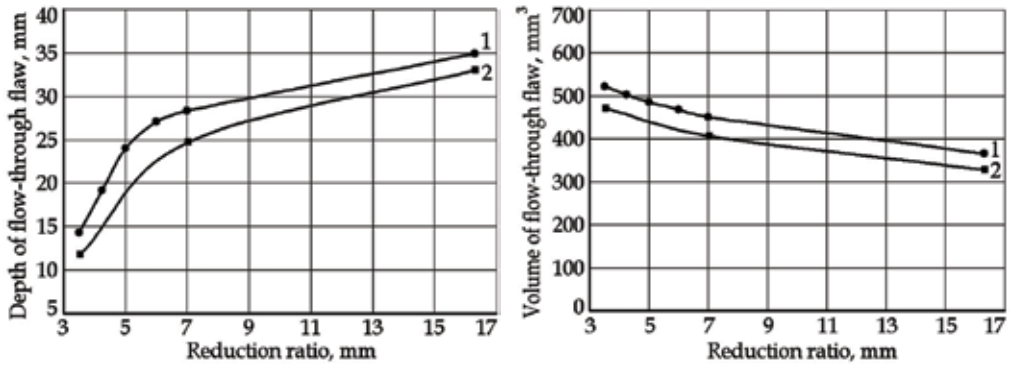
a) b) c) d) e)

Fig. 3.9. The evolution of flow-through flaw: a, b, c, d – are the finite element simulation results ; e – is the photo of upper part of sample with a flow-through flaw: 1 – is the press-washer, 2 – is the porous fibrous pressing

The highest depth of flow-through flaw of 35 mm was reached at extrusion with λ = 16.8 (Fig. 3.10, a) and its volume was about 350 mm^3. The maximal volume of flow-through flawV_{sk} obtained at λ = 3.6 (Fig. 3.10, a) and its depth was minimal, 12-15 mm.
A comparison of theoretical and approximate experimental dependences of depth of flow-through flaw l_{sk} (Fig. 3.10) and height of loosening from the other end of sample h_{raz} (Fig. 3.11) from value of λ has shown that l_{sk} and h_{raz} are significantly growing while increasing of λ that diminishes useful length of sample l^{pr} :

$$l^{pr} = l_0^{pr} - l_{sk} - h_{raz}, \tag{3.2}$$

where l_0^{pr} - is the general length of rod.



a) b)

Fig. 3.10. The maximal depth and volume of flow-through flaw: a- is the dependence $l_{sk}(\lambda)$; b- is the dependence $V_{sk}(\lambda)$: 1- are theoretical dependences; 2- are experimental dependences

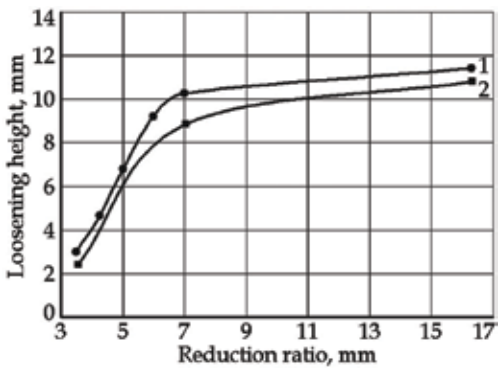


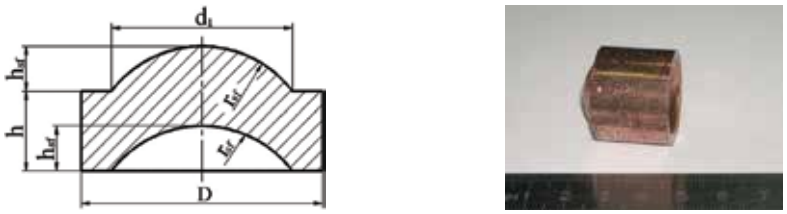
Fig. 3.11. The loosening height: 1 – is the theoretical dependence; 2 – is the experimental dependence

The shape of curves (Fig. 3.10, 3.11) indicates on possibility of their approximation by dependences that are taking into account an influence of non-uniformity of stress-strain state on the volume of flow-through flaw. The effective method of flow-through flaw removal is implementation of billet with compensator (Fig. 3.12). The followings empiric formulas for determination of compensator dimensions have obtained using processing of experimental data by a least-squares method and simulation results:

$$h_{sf} = 2\xi\lambda h_{pr}, \quad r_{sf} = (1.0\text{-}1.7\xi\lambda)D_{pr}, \tag{3.3}$$

where ξ - is the coefficient of non-uniformity of deformation (for copper fibres ξ = 1.02–1.17);
h_{sf} - is the height of compensator;
r_{sf} - is the radius of sphere of compensator;
D_{pr}, h_{pr} - are diameter and height of pressing.

Thus, the stress-strain state at direct extrusion of fibrous pressing is fully determined by reduction ratio. At the reduction ratio λ = 16.8 was produced a compact copper material due to shear stress value exceeding the critical shear stress at high hydrostatic pressure within 1050-1380 MPa that indicates to complete consolidation of fibres. The conditions of temperatures distribution by section of pressing are most hard at λ = 16.8 because of diminishing size of deformation zone and increasing the heat emission to the instrument.



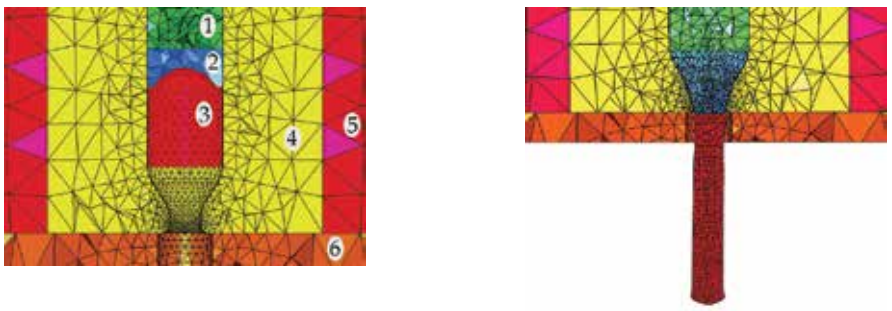
a) b)

Fig. 3.12. The draft of axial section of fibrous pressing with compensator (a) and photo (b)

The dependences for dimensions of defects (flow-through flaw and loosening) in the sample from deforming conditions have been determined. The analytical dependences for dimensions of initial pressing with compensator taking into account a volume of flow-through flaw were obtained and comparing with experimental dependences provided. The results of different methods are corresponding to each other with error less than 10%.

3.4 Modelling of extrusion of porous fibrous pressing with compensator

The investigation of stress-strain state at direct extrusion of porous fibrous pressing with spherical compensator, the reduction ratio λ = 16.8. The finite element model of extrusion of fibrous pressing with compensator is presented on Fig. 3.13. The height of compensator was accepted of 5 mm.



a) b)

Fig. 3.13. The finite element model of extrusion of fibrous pressing with compensator:
a- is the beginning of extrusion; b- is the end of extrusion: 1- is the upper puncheon; 2- is the press-washer; 3- is the initial pressing; 4- is the matrix; 5- is the bandage; 6- is the lower plate

The analysis of distributions of stress intensity and hydrostatic pressure (Fig. 3.14) has shown that type of curves remains analogical to dependences presented on Fig. 3.3. The presence of compensator on pressing provided increasing of stress intensity and hydrostatic pressure in sections 1-1 and 2-2. The hydrostatic pressure in section 3-3 became lower.

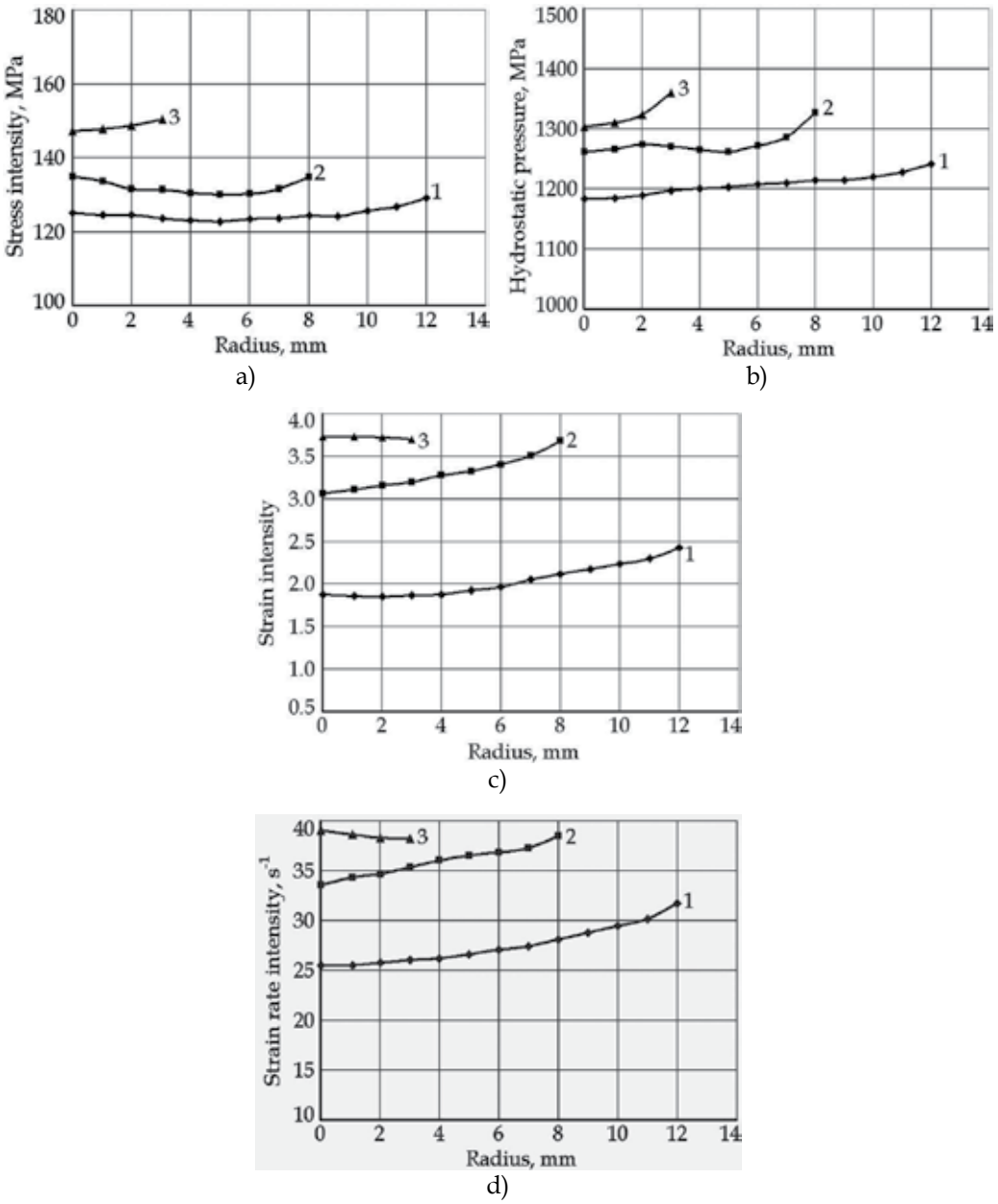


Fig. 3.14. The distribution of stress intensity (a), hydrostatic pressure (b), strain intensity (c), strain rate intensity (d) at extrusion with λ = 16.8: 1– is the section 1-1; 2 - is the section 2-2; 3 - is the section 3-3

There are no inflection points on curves corresponding to beginning formation of flow-through flaw (Shtern et al., 1982; Ryabicheva & Usatyuk, 2006). The distribution of hydrostatic pressure on the section of sample is more uniform. Obviously, the presence of compensator did not exert influence on shear stress intensity. The intensity of deformations is considerably growing in sections 1-1 and 2-2 in the places adjoining to compensator, especially on the axis of pressing (Fig. 3.14, c). The intensity of strain rate was considerably increased too (Fig. 3.14, d).
Thus, the presence of compensator, located on the axis of pressing, resulted to increase of stress intensity and deformations intensity and ensured the removal of flow-through flaw. It has established that deformation takes place more intensively in the area of compensator due to the primary contact of pressing has carried out with press-washer and then with other surface.

3.5 Investigation of plasticity resource

Solving the technological problems of production of fibrous materials coupled with investigation of plasticity resource that is changing under the influence of temperature and strain rate conditions of deformation and is one of criteria for estimation of quality of wares during finite element modelling of direct extrusion of porous fibrous pressings.
The criteria for estimating of plasticity resource were offered on the basis of stress tensor invariants according to (Ogorodnikov et al., 2005). The quantitative relation between ultimate deformation and parameters of stress-strain state is a diagram of plasticity. In such case stiffness of the stress-strain state described by Lode coefficient η_l and exerts influence on plasticity:

$$\eta_l = \frac{\sigma_1 + \sigma_2 + \sigma_3}{\sigma_i}. \tag{3.4}$$

The type of stress-strain state is determined by the Nadai-Lode stress parameter μ_σ that allows estimating an influence of middle main stress on plasticity (Ogorodnikov et al., 2005):

$$\mu_\sigma = \frac{2\sigma_2 - \sigma_1 - \sigma_3}{\sigma_1 - \sigma_3}. \tag{3.5}$$

In such case the measure of plasticity is ultimate deformation that may be determined for any deformed material from a diagram of plasticity built using results of three tests – tension, compression and torsion (Ogorodnikov et al., 2005).
These above mentioned parameters are taking into account of hydrostatic pressure exerting the influence on plasticity, and stress intensity that are determining the plastic flow of material and, also, characterizing the stiffness of stress-strain state. However they are not taking into account the influence of the third invariant of stress tensor.
In the papers (Ogorodnikov et al., 2005; Ogorodnikov et al., 2007) have proposed to construct the diagram of plasticity as a surface of ultimate deformations in space of dimensionless parameters η_l and μ_σ - $e_P(\eta_l, \mu_\sigma)$ for investigation of plasticity resource at the volumetric stress-strain state. During construction of such diagrams the type of loading trajectories and ultimate deformations are simply defined by a deformation scheme and are not depend on properties of material. Therefore, a general view of plasticity resource criterion is presented by following expression (Ogorodnikov et al., 2005):

$$\Psi = \int_0^{e_i} n \frac{e_i^{n-1}}{e_p(\eta_l, \mu_\sigma)^n} de_i \le 1, \tag{3.6}$$

where e_P – is the ultimate deformation at fracture,

e_i – is the intensity of deformations,

$e_P(\eta_l, \mu_\sigma)$ – is the surface of ultimate deformations,

$n = 1 + 0.2\text{arctg}\left(\frac{d\eta_l}{de_i}\right)$ – is the index that takes into account a character of plasticity changing depends on stiffness of stress-strain state.

In the paper (Ogorodnikov et al., 2005) the following dependence was proposed for approximation of surfaces of ultimate deformations:

$$e_P(\eta_l, \mu_\sigma) = \frac{e_P(0,0)\exp(-b\eta_l)}{1 + \lambda_1\mu_\sigma + \lambda_2\mu_\sigma^2}, \tag{3.7}$$

where $\lambda_1 = \ln\left(\frac{e_p(-1,0)}{e_p(0,0)}\right)$, $\lambda_2 = \ln\left(\frac{e_p(0,1)}{e_p(0,0)}\right)$, $b = \lambda_1 - \lambda_2$ - are approximation coefficients;

e_P (0,0) - is the ultimate deformation at torsion test;

e_P (-1,0) - is the ultimate deformation at compression test;

e_P (0,1) - is the ultimate deformation at tension test.

The following values of ultimate deformations were determined by the results of mechanical tests on torsion, compression and tension of material obtained by hot extrusion of fibrous pressing (Fig. 3.15): e_P (0,0) = 0.62; e_P(-1,0) = 0.83; e_P(0,1) = 0.75. The strain rate was 0.1 min^{-1} according to GOST 1497-84.

Construction of surface of ultimate deformations using expression (3.7) makes necessary implementation of strain rate coefficient E_λ that is taking into account the difference in strain rates at mechanical tests and hot extrusion:

$$E_\lambda = \frac{\dot{\varepsilon}_{test}}{\dot{\varepsilon}_{def}}, \tag{3.8}$$

where $\dot{\varepsilon}_{test}$ - is the strain rate at the mechanical tests;

$\dot{\varepsilon}_{def}$ - is the average strain rate in the process of direct extrusion.

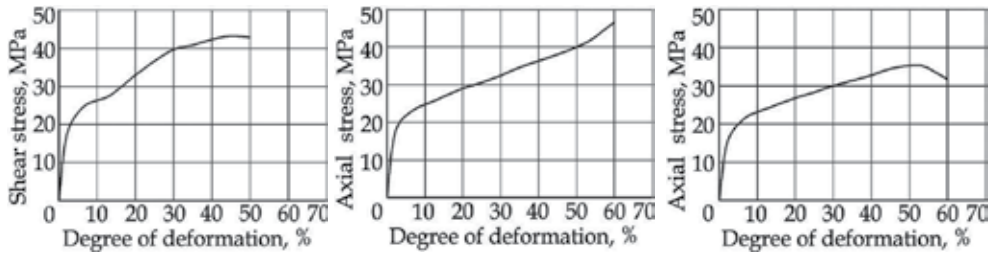
After substitution of formula (3.8) to the expression (3.7) obtained:

$$\lambda_1 = E_\lambda \ln\left(\frac{e_p(-1,0)}{e_p(0,0)}\right), \ \lambda_2 = E_\lambda \ln\left(\frac{e_p(0,1)}{e_p(0,0)}\right). \tag{3.9}$$

Therefore, $\lambda_1 = 1.93 \cdot 10^{-5}$, $\lambda_2 = 1.27 \cdot 10^{-5}$, b = 0.1.

Substituting values from (3.9) to (3.7) having the following expression:

$$e_P(\eta_l, \mu_\sigma) = \frac{0.62\exp(-0.1\eta)}{1 + 1.93 \cdot 10^{-5}\mu_\sigma + 1.27 \cdot 10^{-5}\mu_\sigma^2}. \tag{3.10}$$



a) b) c)

Fig. 3.15. Flow curves: a – at torsion; b – at compression; c – at tension

The dependence of Lode coefficient from intensity of deformations may be obtained by numerical differentiation of (3.4) by e_i for any point of fibrous pressing, if $\eta_l(e_i)$ at interval $[0, e_P]$ continuously differentiable and integrable. This dependence is velocity of changing the stiffness of stress-strain state at hot extrusion and may be decomposed into the trigonometric series that looks like:

$$\frac{d\eta_l}{de_i} = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(e_i) + b_k \sin(e_i), \tag{3.11}$$

where $a_0, a_k = \frac{1}{e_i}\int_0^{e_i} \eta_l(e_i)\cos(k\pi e_i)de_i$, $b_k = \frac{1}{e_i}\int_0^{e_i} \eta_l(e_i)\sin(k\pi e_i)de_i$ – are coefficients.

In the initial moment at $e_i = 0$, $\eta_l(e_i) = 0$ and $\frac{d\eta_l}{de_i} = 0$, therefore $a_0 = 0$ and series (3.11) for the copper porous fibrous pressing may be written in the following way:

$$\frac{d\eta_l}{de_i} = \sum_{k=1}^{\infty} a_k \cos(e_i) + b_k \sin(e_i). \tag{3.12}$$

After substitution of expressions (3.10), (3.11) and (3.12) to (3.6), it looks like:

$$\Psi = \int_0^{e_i} n \frac{e_i^{n-1}}{\left(\frac{0.62\exp(-0.1\eta_l)}{1 + 1.93\cdot 10^{-5}\mu_\sigma + 1.27\cdot 10^{-5}\mu_\sigma^2}\right)^n} de_i \le 1, \tag{3.13}$$

where $n = 1 + 0.2\,\mathrm{arctg}\left(\sum_{k=1}^{\infty} a_k \cos(e_i) + b_k \sin(e_i)\right)$.

Analytical integration of expression (3.13) for obtaining the expression that characterizing a plasticity resource of material at any point of fibrous pressing during passing through the deformation zone is impossible. Numeral integration of expression (3.13) has performed by computer using Mathcad 12. Integration results are presented on Fig. 3.16, curve 1.

Decomposition of function (3.13) in a power-law series have done for saving of computational resources while investigation of plasticity resource in points of fibrous pressing:

$$\Psi = ae_i + \frac{ae_i^2}{2!} - \frac{ae_i^3}{3!} + \frac{ae_i^4}{4!} - \frac{ae_i^5}{5!} + \frac{ae_i^6}{6!} - \frac{ae_i^7}{7!} + \ldots, \tag{3.14}$$

where a – is the coefficient of power-law series.
During extrusion of porous fibrous pressing a = 0.02, then, substituting a in (3.14), calculating factorials and limited to the first seven terms of series having the following expression:

$$\Psi = 0.02e_i + \frac{0.02e_i^2}{2} + \frac{0.02e_i^3}{6} + \frac{0.02e_i^4}{24} + \frac{0.02e_i^5}{120} + \frac{0.02e_i^6}{720} + \frac{0.02e_i^7}{5040}. \quad (3.15)$$

The results of determination of plasticity resource by formula (3.15) are presented at Fig. 3.16, curve 2. The investigation of plasticity resource performed for points located on the axis of pressing while passing through the deformation zone at the reduction ratio $\lambda = 16.8$ shown that for the given deformation conditions the value of $\lambda = 16.8$ is ultimate because of providing the complete consolidation of fibres and exhausting of more then a half of plasticity resource $\Psi = 0.55–0.62 < 1$. Consequently, improving of extrusion productivity by increasing of deforming velocity over 0.5 m/s is not possible. A surface that characterizing intensity of deformations of points of fibrous pressing at hot extrusion $e_i(\eta_l, \mu_\sigma)$ does not intersect the surface of ultimate deformations $e_p(\eta_l, \mu_\sigma)$ (Fig. 3.17) described by expression (3.13).

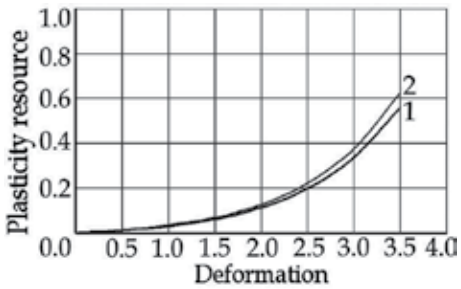


Fig. 3.16. Determination of plasticity resource of points on the axis of fibrous pressing:
1 – obtained by numerical integration of expression (3.13); 2 – according to formula (3.15)

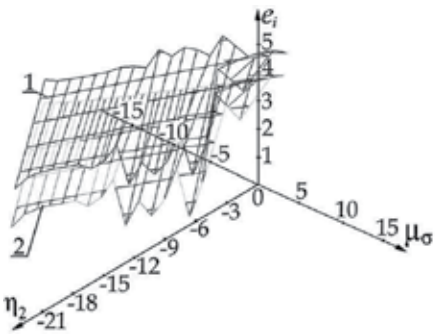


Fig. 3.17. Surfaces of deformations: 1 - is the surface of ultimate deformations; 2 - is the surface of deformations intensity

Comparing values of plasticity resource obtained by numerical integration of expression (3.13) and by using formula (3.15) has shown that they corresponding to each other with relative error 7-11%.

3.6 Modelling of physico-mechanical properties of single-component fibrous material

Finite element modelling of compression test for estimating of physical and mechanical properties of single-component copper fibrous material has been performed.
The initial data are physical and mechanical properties of compact material and value of initial porosity (Table 3.1).

Component	Density, kg/m^3	Initial porosity, %	Young modulus, MPa	Poisson's ratio	Yield stress, MPa	Ultimate stress, MPa
Copper	8940	21	$1.20 \cdot 10^5$	0.33	120	220

Table 3.1. Initial data

The production technique	Ultimate stress, MPa	Yield stress, MPa	Relative elongation, δ,%	Contraction ratio, ψ,%	Hardness, HB
Hot stamping of fibrous pressings, cold deforming, annealing	218.7	45.7	37.5	40.5	55-60

Table 3.2. Mechanical properties of copper fibrous pressing

Material	Kind of data	Porosity, %	Density, kg/m^3	Young modulus, MPa	Poisson's ratio	Relative elongation, δ, %	Yield stress, MPa	Ultimate stress, MPa
Copper	S	5	8490	$1.1 \cdot 10^5$	0.45	38	240	360
	E	3	8670	$1.2 \cdot 10^5$	0.41	40	255	380

S – are simulation results; E – are experimental results.

Table 3.3. Calculated and experimental properties of copper fibrous material after extrusion

3.7 Modelling of physico-mechanical properties of multi-component powder material

Production of antifriction materials with given properties makes necessary investigation the influence of temperature, degree of deformation and strain rate at densification of heterogeneous powder material. The basis of materials observed in this investigation is copper powder obtained from wastes of copper current conductors and ligature is nickel powder produced by recycling of wastes from cadmium-nickel batteries.
The initial data for determination of properties of multi-component copper-based porous powder material are presented in Table 3.4. The finite element model of the multi-component material, the analytical model of compression test and distribution of density are presented on Fig. 3.18, a.
The technology for production of samples consists of the following operations: moulding of powder mixture, sintering at 950 ^{0}C into the synthesis-gas medium for 3.5 hours (the gas composition is 72% H_2, 21% CO, 5.5% CO_2, 1.5% H_2O), repeated moulding up to porosity 10, 20 and 30 %, homogenizing annealing into the synthesis-gas medium at 960 ^{0}C for 1 hour, hardening in water (Ryabicheva et al., 2008).

N	Component	Volume fraction, %	Density, kg/m^3	Young's modulus, MPa	Poisson's ratio	Yield stress, MPa	Ultimate stress, MPa
1	Copper	70-90	8940	$1.20 \cdot 10^5$	0.33	120	220
2	Nickel	10–30	8897	$2.03 \cdot 10^5$	0.31	210	450
3	Cobalt	5	8900	$2.09 \cdot 10^5$	0.31	200	350
4	Iron	2	7850	$2.10 \cdot 10^5$	0.28	200	280
5	Manganese	1	7470	$1.98 \cdot 10^5$	0.22	210	430
6	Titanium	3	4505	$1.10 \cdot 10^5$	0.34	160	530
7	Graphite	1	1800	$0.85 \cdot 10^5$	0.43	100	120
8	Porosity	10–30	0	0.00	1.00	0	0

Table 3.4. The components of multi-component material and their initial properties

The densification process of multi-component powder materials at elevated temperatures is going with shifting of elementary volumes of porous body mainly on phases interface boundaries or «soft» phase. The elements of hard phase are acting like dense bodies. Complex composition of ligature makes an influence on the deforming process. A graphite, for example, does not interacts with copper, remains at free state and may be a hard lubricant on the one part and stress concentrator on the other part diminishing strength and plasticity of antifriction material (Tumilovich et al., 1992; Ryabicheva et al., 2008).

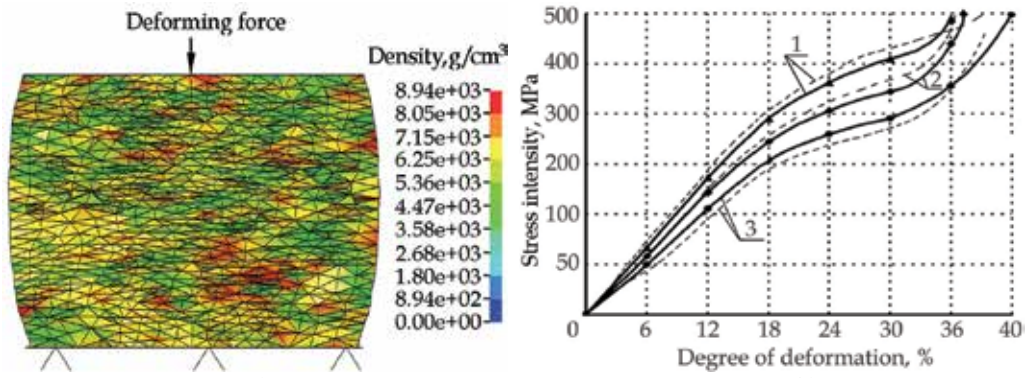
Investigation the influence of degree of deformation and strain rate on densification of heterogeneous powder material at the elevated temperature interval has shown that density is growing the more intensively the higher is strain rate, while increasing the degree of deformation. The most intensive deformation of metal is taking place in deformation zone located in the central part of sample. When stress intensity in hard phase reached the yield stress, the deformation embracing the whole volume of sample. The hardness is higher in zones of higher deformation due to hardening (Ryabicheva et al., 2008).

Metal particles in peripheral ring zone are moving at the radial direction. The shear tensile stresses are arising in it. The hardness is growing while shifting away from periphery of sample that densificating considerably less and is a place of formation of first cracks while reaching the ultimate degree of deformation. The central part of sample is densificating most intensively at the expense of compression stresses and peripheral part less intensively due to metal flow in the radial direction. The condition of reaching the ultimate density is ultimate degree of deformation (Krashchenko & Statsenko, 1981; Ryabicheva et al., 2008).

A transverse flow of metal in the volume of central part have begun after reaching of ultimate degree of deformation and peripheral part is densificating at the expense of central part that becoming smaller. It is impossible to reach full densification in such conditions because of fracturing a surface of sample (Ryabicheva et al., 2008).

It is well-known that higher density of powder material may be reached at higher strain rates and equal degrees of deformation. It has been established experimentally that density of samples is growing up to ultimate while increasing the degree of deformation (Fig. 3.10), and density obtained at strain rate 10 s^{-1} is higher then density of samples upset at strain rate 0.1 s^{-1} at the same degrees of deformation (Ryabicheva et al., 2008).

The initial data for determination of properties of multi-component copper-based porous powder material are presented in Table 3.4. The finite element model of the multi-component material, the analytical model of compression test and distribution of density are presented on Fig. 3.18, a.



a) b)

Fig. 3.18. The finite element model of the multi-component material, analytical model, density distribution (a), stress-strain dependences (b): 1 – is material 1; 2 – is material 2; 3 – is material 3: ▬▬ – are simulation results; - - - - - – are experimental results

Material	Volume fraction, %		Type of data	Porosity, %	Density, kg/m³	Young's modulus, MPa	Poisson's ratio	Ultimate strain, %	Yield stress, MPa	Ultimate stress, MPa
	Copper	Nickel								
Material 1	90	10	S	10	8046	$1.53 \cdot 10^5$	0.42	34	320	430
			E	8	8110	$1.65 \cdot 10^5$	0.40	36	340	460
Material 2	80	20	S	20	7152	$8.75 \cdot 10^4$	0.38	30	280	370
			E	17	7350	$9.15 \cdot 10^4$	0.35	33	300	390
Material 3	70	30	S	30	6560	$5.86 \cdot 10^4$	0.35	30	250	300
			E	32	6245	$5.56 \cdot 10^4$	0.31	28	230	270

S – are simulation results; E – are experimental results.

Table 3.5. Properties of multi-component copper-based powder materials

The laboratory experiments of compression tests are planned and carried out on the basis of numerical simulation results. The stress-strain dependences are drawn using the simulation and experimental results that are in concordance (Fig. 3.18, b). The relative inaccuracy of mathematical and experimental investigation of properties does not exceed 10%.

It has established that highest level of mechanical properties is shown by material 1 due to its lowest porosity. The material 3 has lowest mechanical properties (Table 3.5) because of highest porosity and, moreover, interparticle cracks may appear on copper-nickel boundaries due to significant difference of their strength properties. Thus, decreasing of porosity on 20% promotes to increasing of strength properties on 40%.

4. Conclusion

It has established that stress-strain state and temperature fields at extrusion of fibrous pressing are fully determined by the reduction ratio. A compact copper material was produced at the reduction ratio λ=16.8 and high hydrostatic pressure within 1050-1380 MPa. The shear stress value exceeded the critical shear stress that indicates on complete consolidation of fibres.

Conditions of formation of defects during extrusion of fibrous pressing have determined. The analytical dependences for determining dimensions of initial pressing with a compensator with taking into account dimensions of defects have proposed.

The presence of compensator located on the axis of pressing led to increasing of stress intensity and intensity of deformations and ensured defects' removal. It has established that near compensator deformation is taking place more intensively due to the primary contact of pressing has carried out with press-washer and then with other surface.

Investigation of plasticity resource of points located on the axis of pressing while passing through the deformation zone at the reduction ratio $\lambda = 16.8$ shown that for the given deformation conditions the value of $\lambda = 16.8$ is ultimate because of providing the complete consolidation of fibres and exhausting of more then a half of plasticity resource $\Psi = 0.55\text{-}0.62 < 1$. Consequently, improving of extrusion productivity by increasing of deforming velocity over 0.5 m/s is not possible.

A surface that characterizing intensity of deformations of points of fibrous pressing at hot extrusion $e_i(\eta_l, \mu_\sigma)$ does not intersect the surface of ultimate deformations $e_p(\eta_l, \mu_\sigma)$.

The technique for finite element modelling of physical and mechanical properties of single-component fibrous material with taking into account properties of fibres' material in compact state and deforming conditions that allows defining conditions of complete consolidation of fibres at the deforming process using the stress-strain state analysis results of fibrous pressing at the deforming process has been developed.

The technique for modelling of physical and mechanical properties of multi-component powder materials using a finite element method on the basis of physical and mechanical properties of initial components while accounting deforming conditions has been developed. The distribution of density of multi-component powder material in the volume of sample obtained.

The influence of nickel content and porosity value on mechanical properties of material has been established. Increasing of nickel content leads to enhancing of strength properties. The content of other components and their influence on properties was accounted by interaction of finite elements. Growth of porosity leads to decreasing of mechanical properties. The results of modelling physical and mechanical properties of multi-component powder materials are well concordant with the results of the laboratory experiments.

5. References

Hallquist, J. O. (2006). *LS-DYNA Theory Manual,* Livermore Software Technology Corporation, ISBN 0-9778540-0-0, Livermore.

Kachanov, L. M. (1969). *The Basics of Plasticity Theory*, Nauka, Moscow.

Krashchenko, V. P. & Statsenko, V. E. (1981). Effect of temperature and strain rate on basic processes controlling the strength of copper, *Strength of Materials*, Vol. 13, No. 4, pp. 487-492, ISSN 0039-2316.

Ogorodnikov, V. A.; Kiselev, V. B. & Sivak, I. O. (2005). *Energy. Deformation. Fracture.,* Universum-Vinnitsa, ISBN 966-641-117-2, Vinnitsa.

Ogorodnikov, V. A.; Muzichuk, V. I. & Nahajchuk, O. V. (2007). *The mechanics of cold shaping processes with equitype schemes of deformation mechanism,* Universum-Vinnitsa, ISBN 978-966-641-217-4, Vinnitsa.

Petrosjan, G. L. (1988). Plastic deformation of powder materials, Metallurgy, ISBN 5-229-00160-7, Moscow.

Ryabicheva, L. A. & Usatyuk, D. A. (2006). Using the finite element method for solving of coupled thermal-structural problem, *Herald of the DSEA*, No. 3, pp. 141-147, ISSN: 1993-8322.

Ryabicheva, L. & Usatyuk, D. (2007). Numerical Simulation and Forecasting of Mechanical Properties for Multi-Component Nonferrous Dispersion-Hardened Powder Materials, *Materials Science Forum*, Vols. 534-536, pp. 397-400, ISSN: 1662-9752.

Ryabicheva, L. A. & Nikitin, Yu. N. (2008). Production and properties of copper-based powder antifriction material, *Powder Metallurgy and Metal Ceramics,* Vol. 47, No. 5-6, pp. 299-303, ISSN: 1573-9066.

Ryabicheva, L. A.; Tsyrkin, A. T. & Sklyar, A. P. (2008). Production and properties of copper-based compacts, *Powder Metallurgy and Metal Ceramics,* Vol. 47, No. 7-8, pp. 414-419, ISSN: 1573-9066.

Segal, V. M.; Reznikov, V. I. & Malyshev, V. F. (1981). Variational functional for a porous plastic body, *Powder Metallurgy and Metal Ceramics,* Vol. 20, No. 9, pp. 604-607, ISSN: 1573-9066.

Skorokhod, V. V. (1973). *The rheological basics of sintering theory*, Naukova Dumka, ISBN 5-7695-2294-1, Kiev.

Shtern, M. B.; Serdyuk G. G. & Maximenko L.A. (1982). *Phenomenological Theories of Pressing of Powders,* Naukova Dumka, Kiev.

Tumilovich, M. V.; Kostornov, A. G.; Leonov, A. N.; Sheleg, V. K. & Kaptsevich, V. M. (1992). Porous copper-base fiber-powder materials, *Powder Metallurgy and Metal Ceramics,* Vol. 31, No. 3, pp. 239-242, ISSN: 1573-9066.

Wagoner, R. H. & Chenot, J. L. (2001). *Metal Forming Analysis,* Cambridge University Press, ISBN 0-521-64267-1, Cambridge.

Zienkiewicz, O. C. & Taylor, R. L. (2000). *The Finite Element Method,* Butterworth-Heinemann, ISBN 0-7506-5049-4, Barcelona.

17

Simulation Technology in the Sintering Process of Ceramics

Bin Lin, Feng Liu, Xiaofeng Zhang, Liping Liu, Xueming Zhu
Tianjin University
China

1. Introduction

Ceramics is one of the oldest artificial materials in the world. As a key process of ceramics manufacture, the sintering process, which belongs to the heat engineering technology, can directly influence the quality, yield and cost of ceramic products. Based on the computer, simulation and artificial intelligence technology, the intelligent ceramics sintering can be realized with the research of CAS (Computer-Aided Sintering). CAS technology is a development tendency of the ceramics manufacture combined with heat engineering technology, because with it not only the sintering quality and yield of ceramics products can be improved but also the energy consume can be decreased. Associated with the application of simulation technology, the topics about CAS are discussed as follows:

Basic concept of CAS

Method of search for geometric heat centroidal point (GHCP) using of simulation technology

Simulation temperature field evolution of ceramics body adopting ANN (Artificial Neural Network) technology

Simulative analysis about stress filed of ceramics body

Appropriate processes of ceramics sintering based simulation technology

The ceramic is widely adopted due to its unique and excellent characters. The requirement of the sintering product quality is very high because of its difficult-to-cut character. The factors which influence the quality of the sintering product include not only the roughcast but also the change event of the temperature distribution in roughcast. From another point of view, the factors include the sintering curve. The traditional sintering curve was defined all by the people's experience. The waste of resource is not obvious when the small ceramic product is developed by experimentation. However, the large structure parts like missile spinner fail to sinter once, a huge economic loss will come to being. And from the view of environmental protection and the resources reasonable use, this traditional method is also unsuitable for present industrial development. So, in order to set the sintering curve scientifically, the change event of the temperature distribution in roughcast should be studied and the rule has to be found out (Zeng & Zhang, 1994; Zhao, 1993; Jeong & Auh, 2000). This paper mainly introduces CAS, researches for GHCP on simple shape ceramic

body and complex shape ceramic body using of simulation technology, Simulates temperature field evolution of ceramics body during sintering adopting ANN technology, simulates the stress field of ceramic body during sintering and discusses the appreciate process of ceramic sintering.

2. Important

Neural network has been developed rapidly in recent years. Following the development of large scale integrated circuits and computer technology revolution, complex and time-consuming operation has no longer been the main issue to researchers. So far, dozens of neural network models have been produced which broadly divided into two categories: feed forward network and feedback network. BP algorithm is the most important and common learning algorithm of feed forward network.

Present, neural network has been applied to various fields and achieved very exciting advances in many ways, such as intelligence control, system identification, pattern recognition, computer vision, self-adaptive filtering and signal processing, nonlinear optimization, automatic target recognition, continuous voice recognition, sonar signal processing, knowledge processing, sensing technology, robot technology etc. Neural network has been applied to ceramic industry by more and more scientific and technical personnel recently.

Ming Li etc. use neural network with single hidden layer to simulate the temperature distribution of burner nozzle. In this paper, fuel pressure, atomizing wind pressure and combustion-supporting wind pressure are the input parameters and the average combustion temperature is the output. Intrinsic relationship between the input and output has been set by neural network with single hidden layer which can be fast mapped between them. The network exercised 5770 times by nine sets of data has been tested. The relative error is less than 0.9%, maximum absolute error is 7.44°C. This Indicates that using artificial neural networks to simulate the temperature distribution of burner nozzle is feasible.

Basing on systematic analysis, Guolin Hu, Minhua Luo selected nine identification parameters including the heat insulation time, the average of high temperature section and the heating rate of various stages and built a BP network model to train. 20 samples have been identified using the decided identification model and the accuracy of recognition is 90%.It is shown that the porcelain brick sintering condition can be identified by BP model.

Lingke Zeng, Minhua Luo etc. utilized the mixture ratio and the sintering properties of TZP to train the BP network, and then the performance parameters such as volume density, relative density, linear shrinkage rate of the sintering pattern were predicted. The deviation between the predictive value and the true is very small.

The application of neural network in the ceramic industry is just started, but very successful, especially for the identification, forecast of material properties, analysis and prediction of ceramic material defects and prediction of the dynamic temperature field etc. Further application of neural network in the ceramic industry will be realized. For instance, neural network can be used in temperature field analysis of a ceramic body during the sintering process which is not mentioned in literatures nowadays.

year	author	content
1976	Jinxue Gao	A model of tunnel kiln
1979	D.P.Shelley	Structure design of periodic kiln walls using computer simulation
1981	В.г.Аббакумов	A combustion mathematic model of sintering zone in tunnel kiln
1982	Zhenqun Liu, Lingke Zeng	A tunnel kiln mathematic model based on the calculation of parking stall
1982	Duan Song	Design and operation improvement of tunnel kiln using computer simulation
1993	Lingke Zeng, Gongyuan Zhang	Dynamic measuring of surface temperature field of ceramic body during the firing process
1994	Lingke Zeng, Gongyuan Zhang	3D finite element analysis of temperature and thermal stress fields of ceramic body in sintering course
1997	Chuangliang Chen, Lingke Zeng,	Simulation of periodic kiln walls temperature field
1997	Ming Li	Simulation and study on the temperature distribution of furnace burner using neural network
1998	Guolin Hu, Minhua Luo	Prediction of the porcelain brick sintering condition under various sintering temperature curve using the BP network
2002	Lingke Zeng, Minhua Luo	Prediction of the product performance under different formula and sintering conditions using neural network

Table 1. Research situation of ceramic kilns in recent years

3. Information

3.1 Basic concept of CAS

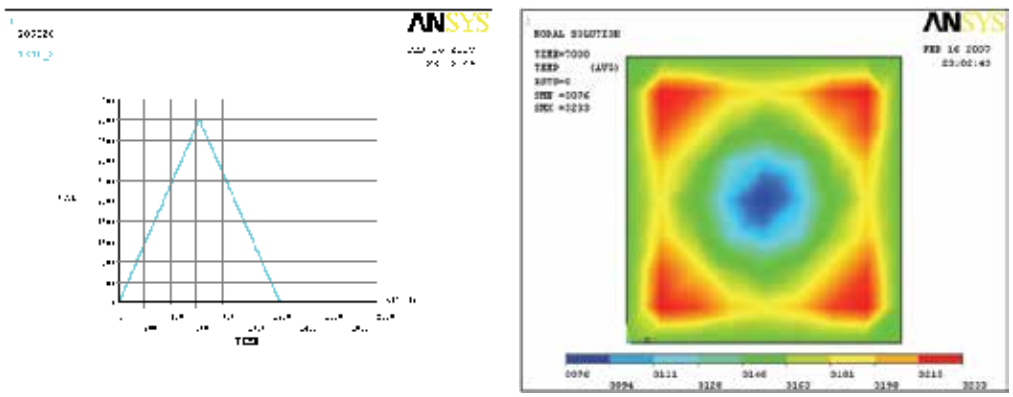
CAS (Computer-Aided Sintering) is used establish of mathematic models of sintering process and simulating this process by computer, finite element analysis and artificial intelligence technology. The temperature and thermal stress distribution fields in the inner of the product under some sintering condition can be required by simulation of the sintering process. So the rational sintering process can be designed to control the temperature and the thermal stress of the sintering process by the simulation results. Naturally the deformation

and cracks during sintering process reduce and the quality of the sintering product improves.

3.2 Method of search for geometric heat centroidal point (GHCP) using of simulation technology

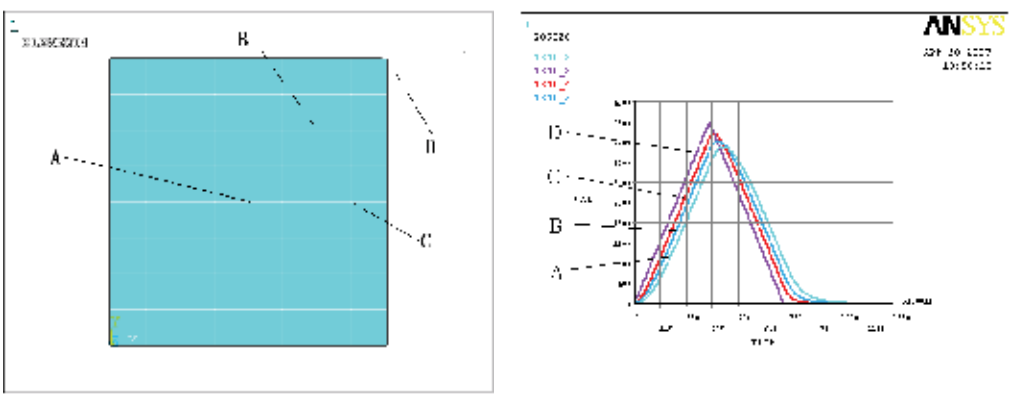
3.2.1 Research for GHCP on simple shape ceramic body

In order to search for geometric heat centroidal point, temperature distribution of ceramic roughcast is analyzed with ANSYS. The shape of the ceramic roughcast is supposed to be square. Temperature load is applied according to the sintering curve (Hong & Hu, 1992). Temperature rise rate k whose unit is °C/s is denoted by the slope angle α of the sintering curve (tanα=k). The 45° sintering curve means that the temperature rise rate is 1°C/s. The initial sintering temperature is 0 °C and the max one is 3600 °C. When reaching the max sintering temperature, the roughcast is cooled according to the same temperature change rate. Taking the 30° sintering curve as example, simulation of the temperature distribution of ceramic sintering with ANSYS is shown as Fig.1.



(a) 30° sintering curve (b) Temperature distribution at 7000s

Fig. 1. Simulation of the temperature distribution of ceramic sintering



(a) The location of the selected no (b) Temperature variation curves of A, B, C, D

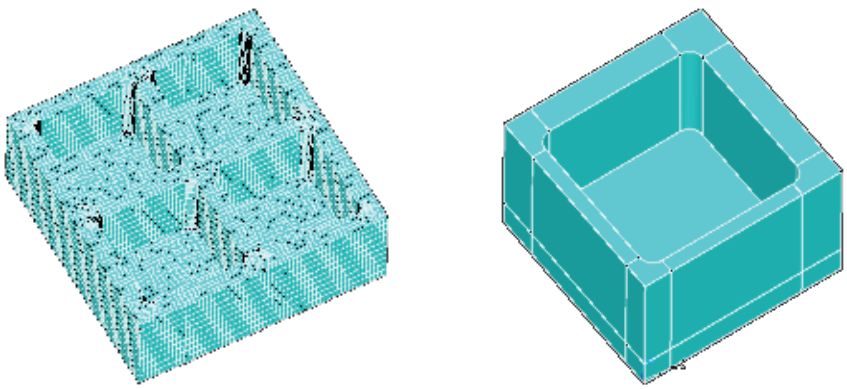
Fig. 2. Four representative nodes and their temperature variation curves

The temperature of every node at each time can be got. Four representative nodes are selected and shown as Fig.2. Temperature difference between node A and D is much larger than that between node B and C. So node A and D whose temperatures are taken into consideration mostly are selected as geometric heat centroidal points of the square ceramic roughcast.

3.2.2 Research for GHCP on complex shape ceramic body

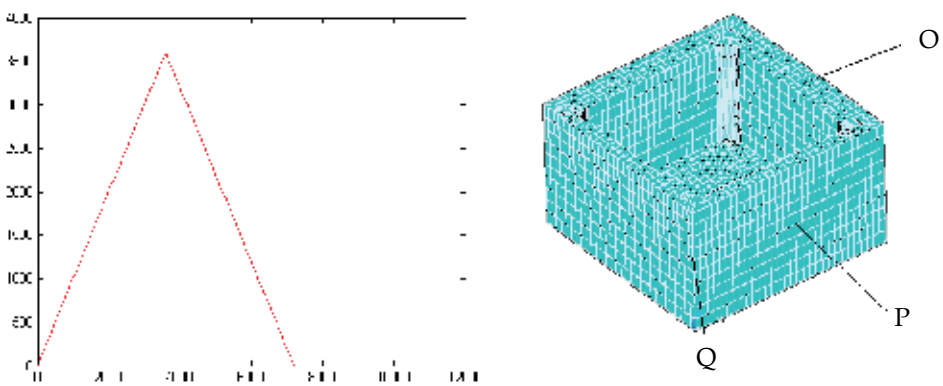
The complex shape ceramic body is shown in Fig.3 (a). This problem belongs to transient thermodynamic issue. Based on its symmetry, a quarter of the ceramic body is used to build a finite element model which is shown in Fig.3 (b).

The temperature load is applied according to the sintering curve whose slope angle is 45° shown in Fig.4(a). Temperature distribution map at different time points are illustrated in Fig5. The value of temperature increases from blue to red. It can be seen from these pictures that the location of geometric heat centroidal point (GHCP) is at notes O, P and Q shown in Fig.4 (b).



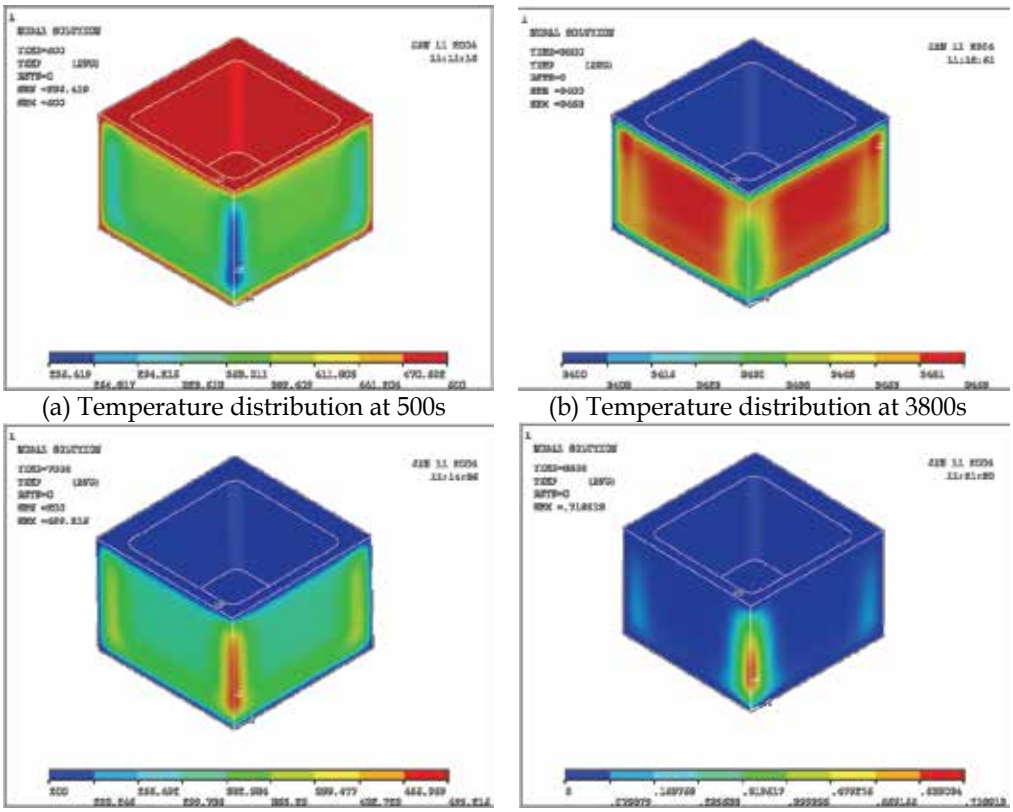
(a) The complex shape ceramic body (b) The finite element model

Fig. 3. The complex shape ceramic body and its finite element model



(a) The 45° sintering curve (b) The location of nodes O, P Q

Fig. 4. The sintering curve and the location of nodes O, P, Q



(a) Temperature distribution at 500s (b) Temperature distribution at 3800s

(c) Temperature distribution at 7000s (d) Temperature distribution at 8500s

Fig. 5. The temperature distribution map

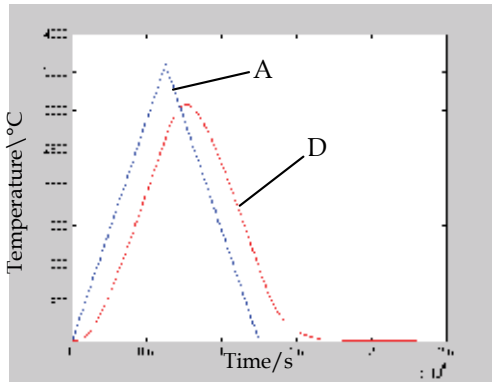
3.3 Simulation of temperature field evolution of ceramics body adopting ANN (Artificial Neural Network) technology

BP network has a strong non-linear mapping ability and a flexible structure. In this paper, a non-linear function f: $y_n x u_{n\times n} \rightarrow$ ý is confirmed to simulate the temperature distribution of ceramic sintering. The following equation having the non-linear mapping relationship is realized by the BP neural network.

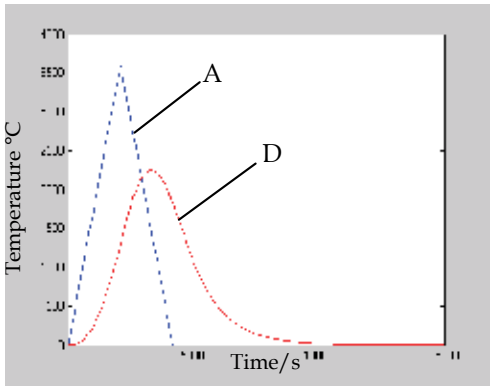
In equation (1), ý is the output of the BP neural network, y is the temperature distribution data of the ceramic GHCP analyzed with ANSYS and also the input of the BP neural network, u is the time series of the input parameter, p is the number of the input parameter. This BP neural network is a series-parallel model.

$$\acute{y}(k+d)=N_f(y(k),\ldots y(k-n+1),u_1(k-1),\ldots u_1(k-n+1),\ldots,u_p(k),\ldots u_p(k-n+1)) \quad (1)$$

The BP neural network is trained by the monitoring way. The input sample of the neural network is very important during training. The result analyzed with ANSYS is used as input sample to train the network in this paper. Ceramic sintering under linear sintering curves with ten different slopes from 5 to 85°has been analyzed with ANSYS. The analyzed data has been used as the training sample of the neural network. The temperature distribution of the ceramic GHCP A and D analyzed with ANSYS is shown as Fig.6.



(a) Temperature distribution of the ceramic HGCP A and D analyzed under 30°sintering curves



(b) Temperature distribution of the ceramic HGCP A and D analyzed under 60°sintering curves

Fig. 6. Input sample of the BP neural network

During training BP neural network, there usually happens platform phenomenon, which is false saturation and makes BP neural network constringe slowly. The reason of appearance of platform phenomenon is: When all of the neuron input attains saturation area, the derivative of the saturated non-linear neuron function approaches zero, which causes weight and valve value can not update effectively. For the sake of reduction or elimination of the Platform phenomenon, neural network has been analyzed and adjusted according to following several aspects (Li, 1996; Xie & Yin, 2003).

The sample value is normalized into range from 0.1 to 0.9 by equation (2). Where x_i is normalized sample value. x_{min} and x_{max} express the minimum and maximum value of x_i, respectively.

$$x_i = \frac{0.8}{x_{max} - x_{min}} \cdot x_i + \frac{0.1x_{max} - 0.9x_{min}}{x_{max} - x_{min}} \tag{2}$$

The preliminary weight value is set up randomly in the training process of the BP neural network. In order to rapidly constringe of the neural network training process and reduce Platform phenomenon, the preliminary exciting value is selected within ±0.01 in this paper.

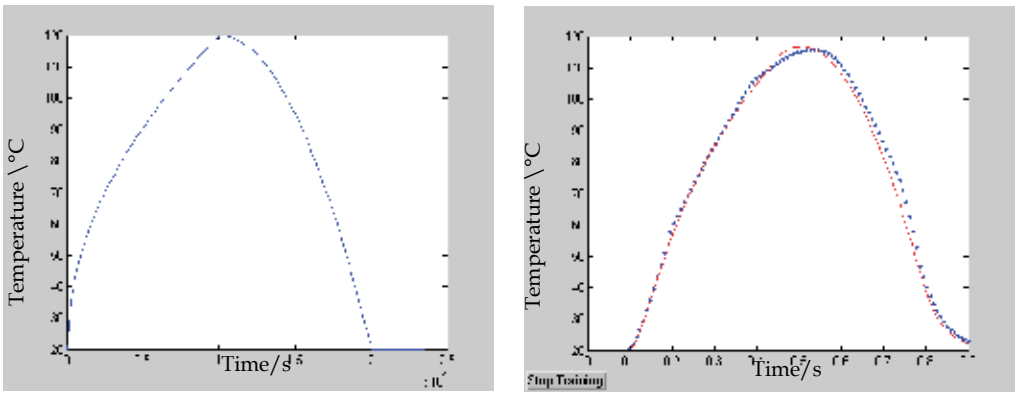
Sigmoid function including logarithm function, hyperbolic-tangent function and so on is adopted widely in BP neural network. In this BP neural network, hyperbolic-tangent function is used as the neuron function in hidden layer, and the linear function is used as neuron function in out-put layer.

The topology of the entire neural network plays a key role. The node number of the input layer and the output layer is easily ascertained by the number of input parameter and output parameter. Thus, the neuron number of the hidden layer is the key to determine the topology of the neural network. If the neuron number of the hidden layer is too small, it will seriously affect the approximation ability of the neural network. If the neuron number of the hidden layer is excessive, it will aggravate the burthen of the neural network. The neuron number of the hidden layer is selected 80 in this study.

Dynamic study rate η is adopted to accelerate the BP neural network convergence. The dynamic coefficient mc make the weight value use the trained information. In training

process, the weight value varies toward the last adjusted result. Selecting optimum study rate η and dynamic efficient mc will accelerate BP neural network convergence and decrease platform phenomenon. When the study rate is 0.075, the neural network converges fastest, and the training time is least. The bigger the dynamic efficient mc is, the higher the convergence speed of the neural network is. If the dynamic efficient mc is too big, it will make the convergence of the neural network unsteady and the kinds of instable factors will increase, too. As a result, the local convergence usually happens in training network. When the trained results differ little at different dynamic efficient, the smaller dynamic efficient mc is selected.

The trained neural network is tested by the sample analyzed with ANSYS under non-linear sintering curve. The input sample of the test and the tested result is shown as Fig.7. The biggest error is within 5°C. So the temperature difference of the ceramic HGCP can be forecasted fast by the trained BP neural network (Liu et al., 2010).



(a) Testing input sample (b) Tested result

Fig. 7. Testing the BP neural network

3.4 Simulative analysis about stress filed of ceramics body

Temperature (°C)	Density (kg/m³)	Specific heat (J/kg °C)	Thermal conductivity (W/m°C)
<900	1800-0.22T	836.8+0.263T	0.71+1.03*10⁻³T
900～1200	382.5+1.355T	836.8+0.263T	0.88+1.22*10⁻³T

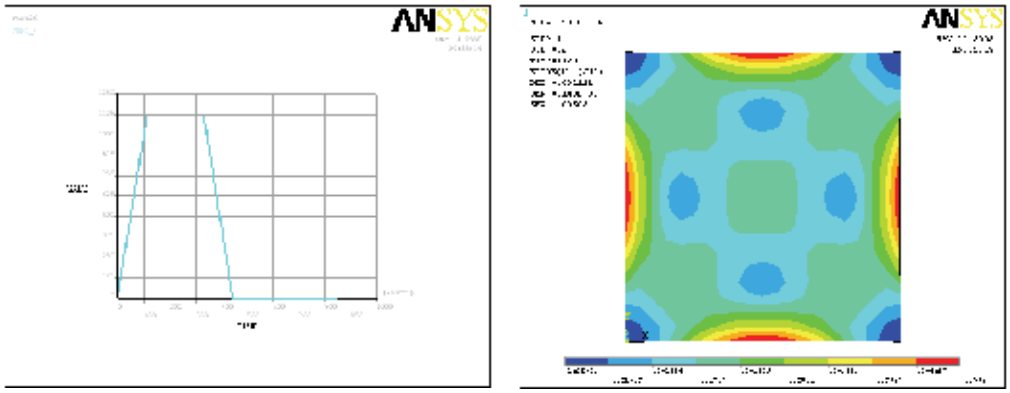
Table 2. Material properties of ceramic

Elastic modulus E(Gpa)	Poisson's ratio μ	Linear expansion coefficient α_1 (m/m°C)
200	0.3	1.3×10⁻⁶

Table 3. Material properties of ceramic

The shape and model of the ceramic body are described at 3.2. The values of the thermal conductivity, specific heat and density are shown in Tab.2, and the elastic modulus, poisson's ratio, linear expansion coefficient shown in Tab.3.

3.4.1 Stress analysis of the traditional sintering curve



(a) Traditional 45°sintering curve (b) Stress distribution map at 1120s

Fig. 8. Simulation of the stress distribution of ceramic sintering

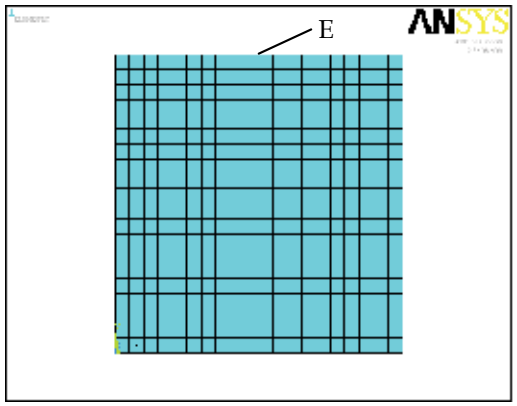


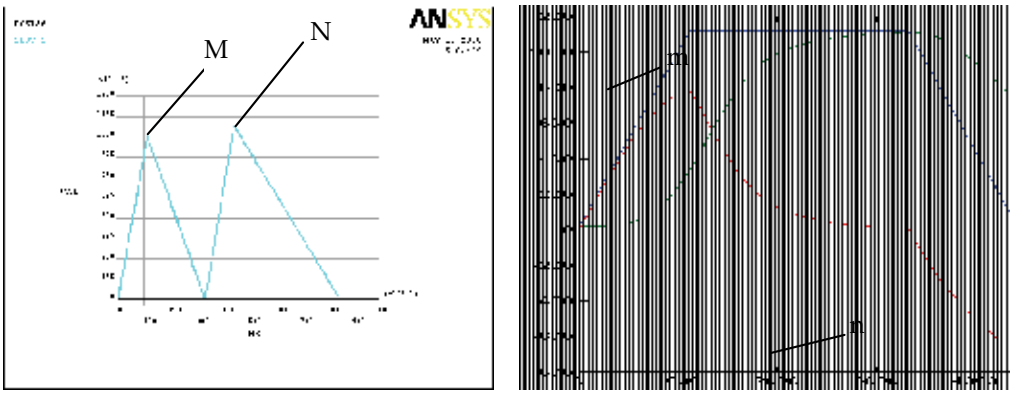
Fig. 9. The location of node E

Since the tensile stress is the main reason of product destruction during ceramic sintering, the first principal stress is elected as the basis for analysis. The temperature load is applied according to the sintering curve whose slope angle is 45° shown in Fig.8 (a). when the outside body temperature rises to 1120 degrees, The stress distribution is illustrated in Fig. 8 (b). The maximal stress value appears at node E which is not the maximal temperature difference node A and D. The node E is illustrated in Fig.9. The stress change at node E during the whole sintering process is illustrated in Fig.10 (a). The maximum stress at node E appears twice respectively at 1120s and 4440s which are exactly the two time points of the maximum temperature difference. When temperature distribution is uneven the thermal stress appears in older to maintain the continuity of displacement. It is shown that the basic cause of thermal stress is temperature variation.

The maximal tensile stress is 0.975165E09Pa at 1120s which is the finish time of heating and also the start time of the first temperature holding, and 0.104123E10Pa at 8440s which is the finish time of cooling and also the start time of the second temperature holding. This indicates that more temperature variation during heating or cooling will cause larger

temperature difference between A node and D node, and then the holding make the temperature difference tend to be uniform. It is shown that the change process of stress illustrated in Fig.10 (a) is firstly from zero to the peak in heating time, from the the peak to zero in the first holding time, secondly from zero to the peak in cooling time, from the the peak to zero in the second holding time. The two peak pressure points are points M and N, respectively corresponding to m and n in Fig.10 (b)

The higher the temperature difference the higher the stress. The more the alternate changing times of extreme pressure the poorer ceramic quality Cracks. All that causes deformation and other defects at node E.

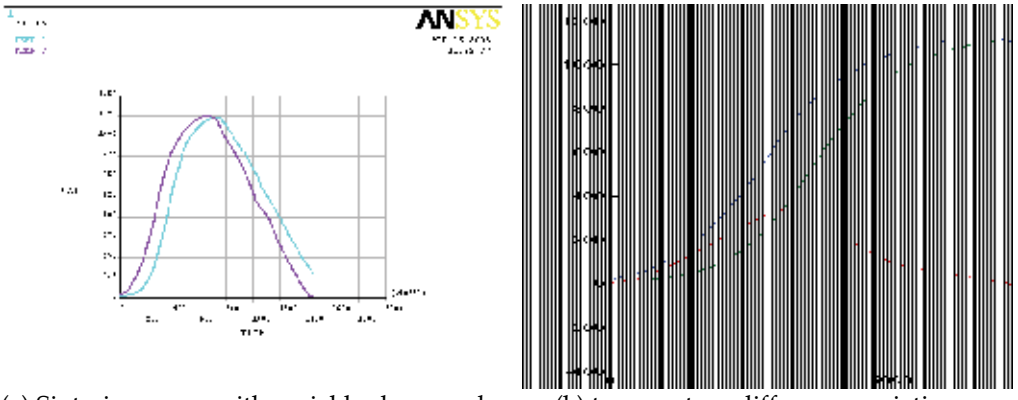


(a) The stress curve (b) The temperature difference curve

Fig. 10. The stress curve and the temperature difference curve

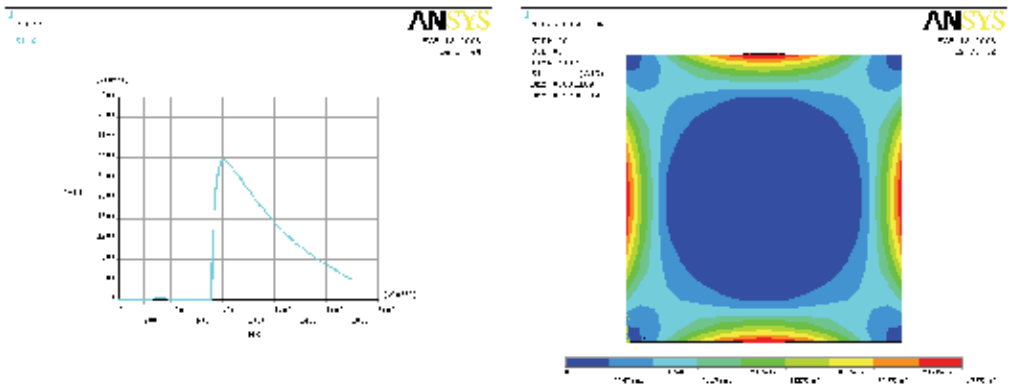
3.4.2 Stress analysis of variable slope curve

The temperature load is applied according to the sintering curve whose slope angle is variable shown in Fig11 (a). The temperature difference variation curve between node A and node D gotten after thermal analysis by indirect method is illustrated in Fig.11 (b). The stress distribution map at 8109s and the stress variation curve at node E during the whole sintering process are shown in Fig.12.



(a) Sintering curve with variable slope angle (b) temperature difference variation curve

Fig. 11. The sintering curve and The temperature difference curve



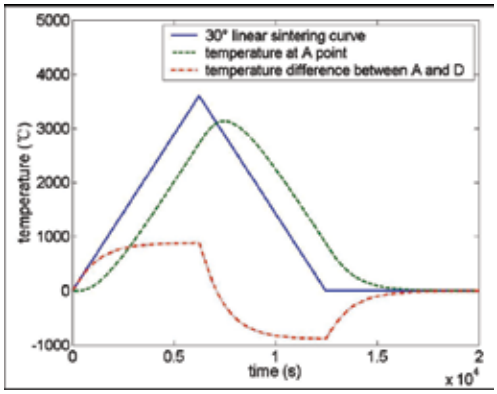
(a)The stress variation curve at node E (b)The stress distribution map at 8109s

Fig. 12. The stress distribution map at 8109s and variation curve at node E

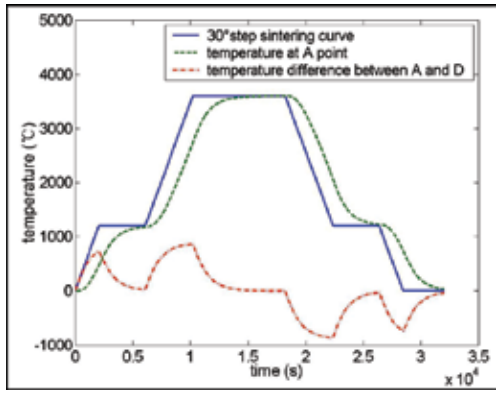
It can be seen from the charts that there is not significant temperature insulation process and temperature difference changes slightly. During the whole sintering process only one pressure peak whose value is 0.278387E+09Pa appears at 8109s during cooling at node E.
Sintering curve with variable slope being adopted, the maximum stress is 26.7% of conventional sintering curve, however, the time expended is 95.5%. During the whole sintering process, the pressure peak appears only once during cooling when the ceramic body is still in the plastic deformation stage. So the damage caused by stress is very small.
The conclusions can be drawn from the above analysis: for simple symmetrical ceramic body, adopting variable slope sintering curve is more reasonable, safer and more effective than the traditional fixed-slope curve.

3.5 Appropriate processes of ceramics sintering based simulation technology

There is an appropriate processes during ceramic sintering. Temperature variation of GHCP under different sintering process reveals this mystery. The temperature variation curves of node A and D under both the linear firing curves and step firing curves with slope angles of 30°, 45°, 60° are shown in Fig.13~15. The temperature difference curves between node A and D are shown in Fig.13~15, too (Zhang et al., 2008).
The max value appears at the second wave crest of the temperature difference curve in firing process under the step sintering curve in Fig.13~15. In Fig.13, the heat preservation is applied at the time of the temperature difference curve approaching the platform area. By now the reduction of the max temperature difference is very small, only 2.06%. In Fig.14, the heat preservation is applied at the time of the temperature difference curve just leaving the overlap area. The reduction of the max temperature difference increases slightly, about 8.42%. In Fig.15, the heat preservation is applied at the time of the temperature difference curve being at the overlap area. The reduction of the max temperature difference achieves about 17.3%.
The result indicates that: the max temperature difference can not be reduced effectively by joining the heat preservation process at any time; the max temperature difference can be reduced effectively when the heat preservation process is applied at the time of the temperature difference curve being at the overlap area; the effect is worse when the curve is near the platform area. So it is necessary to analyze the temperature difference curve for choosing the heat preservation time properly.

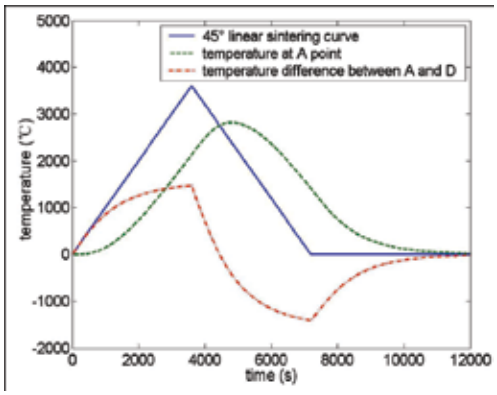


(a) 30° linear sintering curve

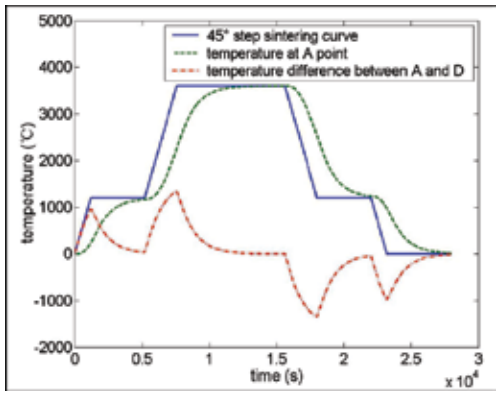


(b) 30° step sintering curve

Fig. 13. The effect comparison of 30° linear sintering curve and step sintering curve

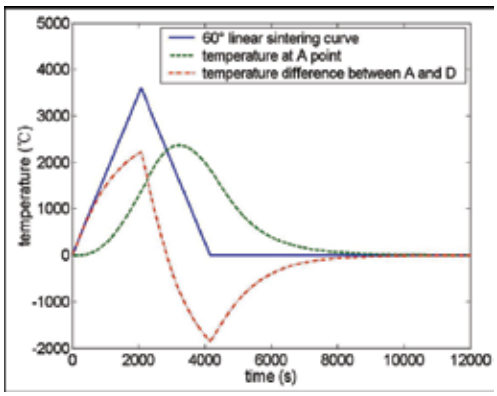


(a) 45° linear sintering curve

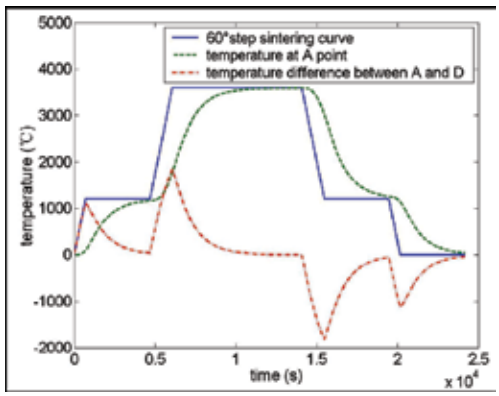


(b) 45° step sintering curve

Fig. 14. The effect comparison of 45° linear sintering curve and step sintering curve



(a) 60° linear sintering curve



(b) 60° step sintering curve

Fig. 15. The effect comparison of 60° linear sintering curve and step sintering curve

Slope angle of sintering curve	Linear sintering curve	Step sintering curve	Reduction
30°	883.6 °C	865.43 °C	2.06%
45°	1472.3 °C	1348.3 °C	8.42%
60°	2220.9 °C	1836.6 °C	17.3%

Table 4. The max temperature difference in ceramic roughcast under different sintering curves

4. Conclusion

1. The trained BP neural network has certain precision and can be used to simulate the changing temperature distribution of the ceramic sintering.
2. The temperature difference of the ceramic HGCP can be forecasted fast by the trained BP neural network. The forecasted results can be used to precisely control the process of the ceramic sintering.
3. The slope of temperature difference curve changes from a max value to zero. When the slope of the firing curve increases, the max temperature difference increases very fast. There are overlap area and platform area in all the temperature difference curves. All the temperature difference curves change from overlap area to platform area.
4. The max temperature difference can not be reduced effectively by joining the heat preservation process at any time. The max temperature difference can be reduced effectively by applying the heat preservation process at the time of the temperature difference curve being at the overlap area. The effect is worse when the temperature difference curve approaches the platform area. It is necessary to analyze the temperature difference curve for choosing the heat preservation time properly.

5. References

Hong, Y; Hu, X. L. (1992). Heat Transfer Mathematical Model within a Ceramic Body during the Firing Process. *China Ceramic*, Vol.28, 1-5

Jeong, J. H; Auh, K. H. (2000). Finite-element simulation of ceramic drying processes. *Modelling Simul.Sci.Eng*, Vol.8, 541-556

Li, J. C. (1996). Application and Realization of Neural Network. Publishing house of University of Electronic Science and Technology, Xi`an

Liu, L. P; Lin, B.& Chen, S. H. (2010). Simulation of Temperature Distribution During Ceramic Sintering Based on BP Neural Network Technology. *Advanced Materials Research,* Vols. 105-106, 823-826

Xie, Q. H.; and Yin, J. (2003). Application Neural Network Method in Mechanical Engineering. Publishing house of Mechanical Industry, Bei Jing, China

Zhang, X. F.; Lin, B & Chen, S. H. (2008). Numerical Simulation of Temperature Distribution During Ceramic Sintering. *Applied Mechanics and Materials* Vols.10-12, 331-335

Zhao, X. L. (1993). Study on the Best Fire Curve of Ceramic Tile and Brick. *Ceramics*, Vol.102, 3-11

Zeng, L. K.; Zhang, G. Y. (1994). Three Dimensional FEM Study of Temperature Field of Ceramic Bodies in the course of firing. *Chinese Journal of Materials Research*, Vol.8, No.3, 245-252

18

Numerical and Experimental Investigation of Two-phase Plasma Jet during Deposition of Coatings

Viktorija Grigaitiene, Romualdas Kezelis and Vitas Valincius
Lithuanian Energy Institute, Plasma Processing Laboratory, Breslaujos str. 3, Kaunas Lithuania

1. Introduction

Atmospheric pressure plasma spraying is widely used to produce various coatings, especially hard ceramic coatings for wear and corrosion protection and thermal barrier function, porous catalytic coatings for environment control and protection, hydrophobic coatings, etc. The plasma spraying process uses a DC electric arc to generate a jet of high temperature ionized plasma gas, which acts as the spraying heat source. The sprayed material, in powder form, is carried into the plasma jet where it is heated, partially or fully melted and propelled towards the substrate. The properties of the produced coating are dependent on the feedstock material, the thermal spray process and application parameters, and post treatment of the coating. However, the influence of flow and particle temperature and velocity on coatings characteristics, its adherence to the substrate, reproducibility of its properties and quality is not clearly established [Fouchais et al., 2006]. Generally, to correlate coating properties to flow parameters and particle in-flight characteristics experimental procedure is used. To monitoring the whole plasma spraying process (plasma jet generation, powder injection, formation of the coating) same techniques, as plasma computer tomography (PCT), particle shape imaging (PST), particle flux imaging (PFI) [Landes, 2006] are used. Such techniques are expensive and complicate for use in industry. Numerical investigations of plasma spray process generally is focused on investigation of heat transfer between plasma jet and surface [Garbero et al., 2006], substrate temperature influence on coatings morphology, adhesion, chemical processes between substrate material and deposited material [Yeh, 2006, Kersten et al., 2001].
In this paper, by means of Jets&Poudres software [Delluc et al., 2003], a numerical simulation of interaction of plasma jet and dispersed particles was investigated. Simulation results were compared with experimental data.

2. Methodology

Numerical research of two-phase high temperature jet was carried out using "Jets&Poudres" software [Delluc et al., 2003], created on the basis of General Mixing (Genmix) software improved by using thermodynamic and transport properties closely related to the local temperature and composition of the plasma. For a particle in a plasma

jet, two characteristics are studied: motion (trajectory, velocity) and thermal evolution (temperature, physical state, heat flux). Thermodynamic and transport properties of the gases are obtained from the T&TWinner database [http://ttwinner.free.fr]. The coating material particle characteristics are also available as a data base. Calculations are carried out for air plasma at atmospheric pressure flowing from jet reactor exhaust nozzle to substratum. When the parameters of plasma jet are achieved as desirable, hard spherical dispersed particles are injected into the flow. Performing modeling and calculating the deformations of the plasma jet thermo fields are disregarded, inlet profiles of temperature and velocity are rectangular shaped and correspond to estimated experimental data [Kezelis et al., 1996]. Plasma jet flows in one direction and the flow is stable, without recirculation and diffusion effects. The numerical simulation results were compared with experimental data.
Experimental plasma spraying system [Valincius et al., 2003] consists of linear DC plasma generator (PG) 30 – 40 kW of power with hot cathode and step-formed anode, plasmachemical reactor, systems of power supply and regulation, PG cooling, feeding and dosing. The operational characteristics of plasma generator are represented in [Valincius et al., 2004].

Regime	I	II	III
P, kW	49	49	49
G, gs^{-1}	5,5	5,5	5,5
$G(H_2)$, gs^{-1}	0	0.1	0,15
T, K	2700	3400	3770
X, mm	70	70	70
V, m/s	1000	1400	1580

Table 1. Plasma spraying regimes for Al_2O_3 films deposition

During plasma spraying experiments the operating conditions of plasma torch were maintained constant. The capacity of plasma torch, total mass flow of air, cooling water and it temperature were measured and from this data plasma jet temperature calculated (see Table 1.). Injection of hydrogen was used to vary outlet plasma jet temperature and velocity, while plasma torch parameter was stable. Powder injection was provided into reactor, which was connected directly to plasma torch anode. Micrographs of the Al_2O_3 powder and sprayed films morphologies were collected using a scanning electron microscope and optical microscope. The spayed particles were collected into distilled water.
These granules can be industrially used as high temperature insulating material. Other primary data (determined by experiments) are as follows: flow outlet nozzle diameter d = 10^{-2} m; the diameter of particles 50 - 70 μm; the exhaust jet is surrounded by air of unrestricted space. The computing domain is a cylinder-shaped space covered with a set of meshes of a grid. The diameter of the computing domain is 200 mm and the total number of variable size geometrical grids is approximately 300000. This is described in detail in [Valinciute, 2007].

3. Results

After mixing with plasma jet, solid particles need some time to heat and at the start their temperature is lower than the temperature of plasma gas. Particles are small-sized and quickly heat up; they are heated in plasma jet by convection, whereas inside particles the

heat is transferred by conduction. As it can be seen from Fig. 1(a), the temperature of dispersed particles near substratum surface exceeds average temperature of gas jet and is 1200 – 1600 K.

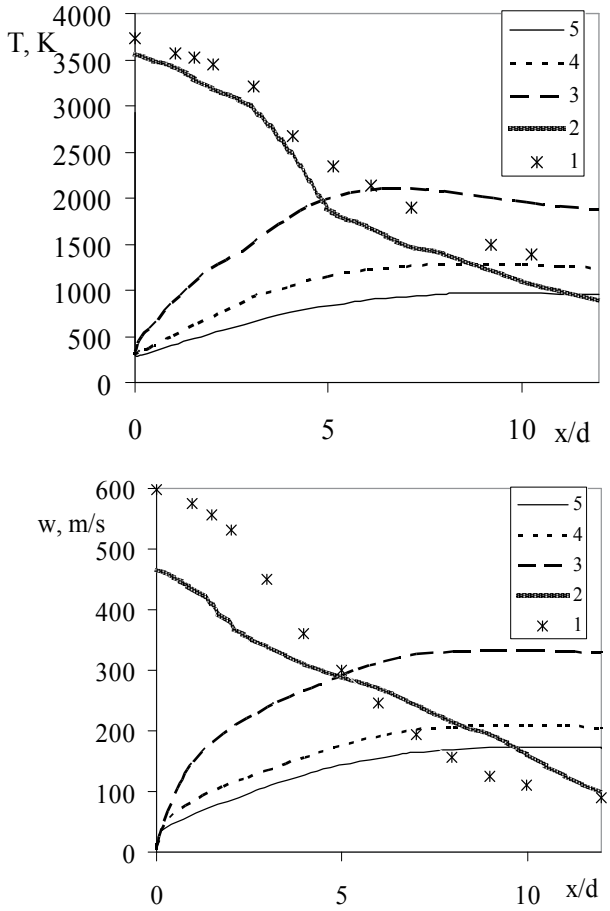


Fig. 1. Distribution of temperatures (a) and velocities (b) of Al_2O_3 particles and plasma jet determined by measurements along the spraying distance. 1, 2 show plasma jet experimental and numerical simulation results respectively, 3, 4 and 5 represent particles of 75, 50 and 35 μm in diameter respectively. x/d is a dimensionless distance

As can be seen from Fig. 1b, velocity of dispersed particles near the covering surface exceeds average gas jet velocity and depending on the sizes of particle reaches 150 – 320 $m \cdot s^{-1}$. The smallest particles achieve higher speed than bigger ones, so, the deciding factor of velocity changes is a resistance force. The velocity of particles stabilizes at x/d = 7 and then the size of particles almost has no significant influence. The surface of substratum at the distance x/d = 8 – 12 will be hit stable force by the jet stream and the value of kinetic energy is ultimate. Figure 2 represents the proportional distribution of plasma jet and dispersed ceramic particles temperatures, measured or calculated by different authors [Delluc et al.,

2005, Klocker et al., 2001]. The trajectories of plasma flow are very similar and have a near agreement. Some differences at the end of travel distance can be observed. Disagreement occurs due to different experimental set-up operating conditions, numerical simulation options, and plasma spraying process regimes.

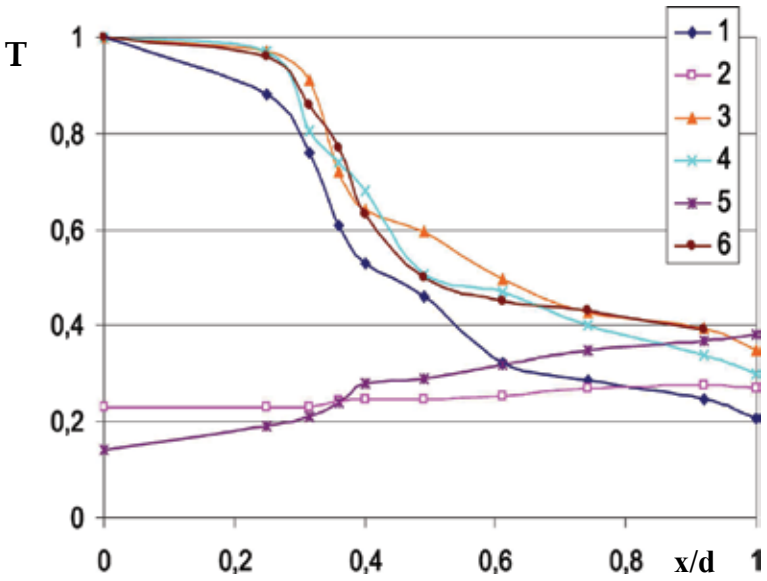


Fig. 2. Nondimensional distributions of plasma temperature (*1* - calculated with "Jets&Poudres" by other authors [Delluc et al., 2005], *3* - our experimental research, *4* - calculated with "Jets&Poudres", *6* - calculated by other authors using other numerical models [Klocker et al., 2001]) and ceramic 50 μm particles' temperature (*2* calculated with "Jets&Poudres" by other authors [Delluc et al., 2005], *5* - our calculation with "Jets&Poudres")

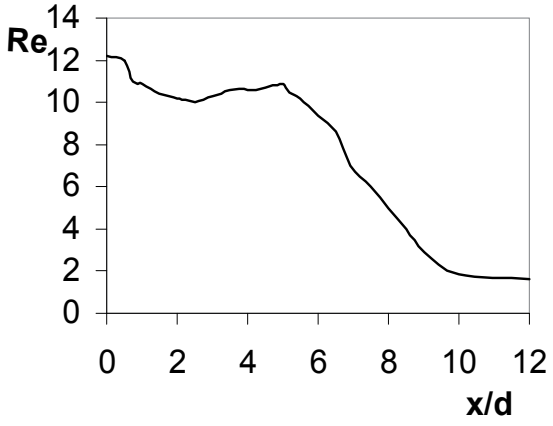


Fig. 3. Variation of Reynolds number along spraying distance

Variation of curve Reynolds number (Re) along flow axis is presented in Fig. 3. In our case, for the regime *I* in Table 1 the value of Re varies from 2 to 12. The largest value of Re is found near the outlet. Since jet mixes with the ambient air and is interrupted, flow becomes

unstable. On further gas in the jet cools down and slightly stabilizes itself. At a distance x/d = 3 from exhaust nozzle, Re value slightly increases since in this period the jet is slightly disturbed. At this moment a very intensive melting of particles occurs and recirculation zone appears. At x/d = 8 – 9 from exhaust nozzle a particle does not melt and flow stabilizes, whereas Re number obtains a steady value.

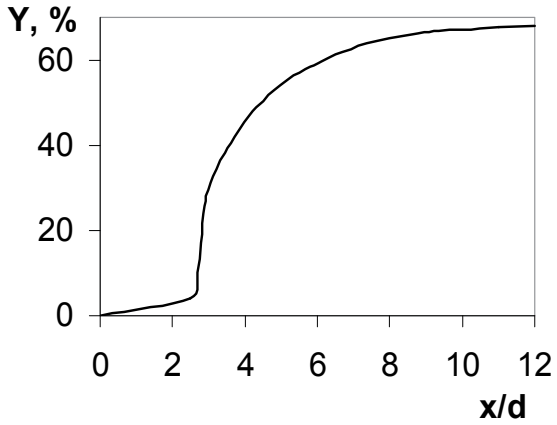
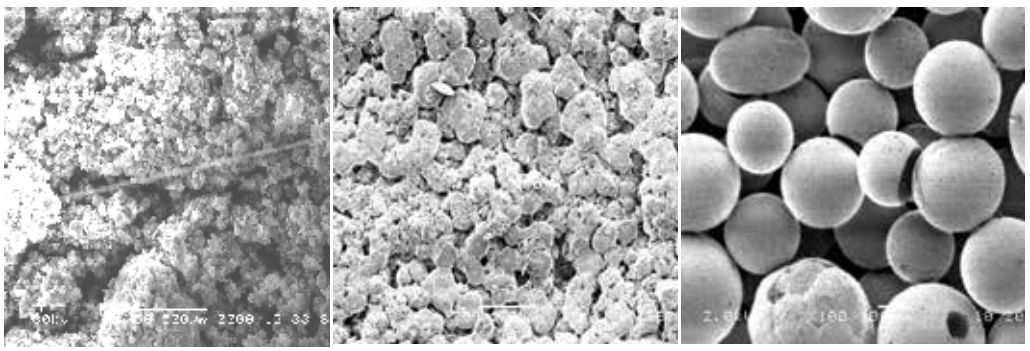


Fig. 4. Dependence of melting degree of 50 μm Al_2O_3 particle from spraying distance

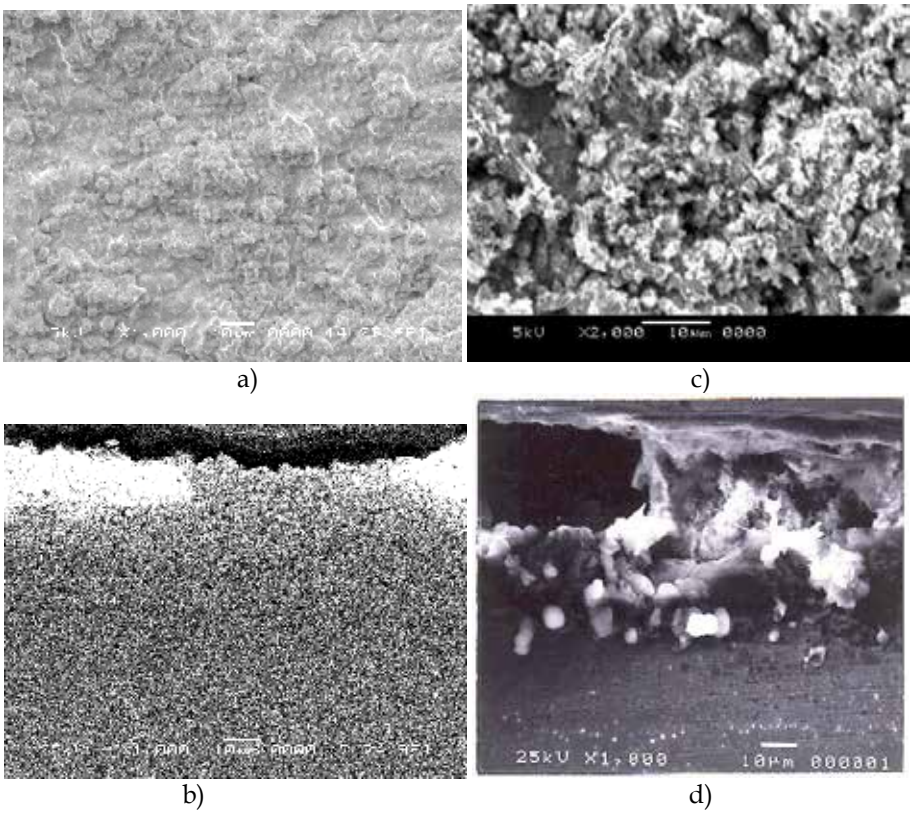


a b c

Fig. 5. SEM micrographs of initial powder (a) and after passing through the plasma jet: (b) at x/d = 35-40 mm outlet form nozzle, (c) the granules produced at x/d = 10 from outlet nozzle

Intensity of particle's melting (*Y, %*) in jet depending on travel distance along the flow axis is presented in Fig. 4. The interaction between high temperature jet and injected particles begins immediately. The particle, injected into plasma jet, passes three main flow zones until it reaches a fixed substratum: heating of the particle, its melting, and stable flow. As can be seen from results, initial heating period of the particle continues to x/d = 2.7 – 3. During this time the largest part of plasma energy is used for heating the particle. When particle is heated up, it begins to melt due to physical and chemical conversions inside it. Temperature of particle gradually rises and melting rapidly proceeds. The most rapid melting occurs at distance x/d = 3 – 8 from exhaust nozzle and this is the second melting zone of particle. The practical usability of calculation results has been verified by comparing the simulation data with experiments

[Valatkevicius et al., 2003, Brinkiene et al., 2005]. Morphologies of plasma-sprayed Al_2O_3 powders during the *II* regime (Table 1) are shown in Fig. 5. As observed by scanning electron microscopy, the initial powder is in the form of agglomerates with wide size distribution. To determine the melting degree, shape, and size of sprayed particles, they have been collected into distilled water at different distances from outlet nozzle. After passing x/d = 3.5 - 4, the particles appear partially melted (Fig. 5(b)). During the melting of initial particles of 100 μm in diameter the plasma spray pyrolysis process occurred. Dispersed particles of Al_2O_3 injected into arc column showed a very fast bulk melting and then very fast particle surface cooling. Further from plasma torch nozzle to the substratum the particles turn into very large granules with the diameter of 150 - 200 μm (Fig. 5(c)). When the coatings are produced, particles resolve into small fragments on their way and splash on the surface of substratum. Sharp edges of particles become round and the surface of coating becomes fine and smooth (Fig. 6). Applying the *I* regime of plasma generator (see Table 1) and regulating the working gas flow, PG arc current, spray distance, and at initial diameter of 30 - 50 μm of dispersed particles, the porous coatings with large free surface for catalytic application (Fig. 6(c, d)) are obtained. Applying the *III* regime, dense thin films for protective purposes could be deposited (Fig. 6(a, b)). In the latter case the plasma spray pyrolysis effect has occurred and initial dispersed particles have broken up into a large amount of fragments. Consequently the grains of plasma sprayed coatings were smaller than 5 μm.



a) c)

b) d)

Fig. 6. SEM micrographs of dense and porous plasma sprayed alumina coatings: (a, c) surface morphology, (b, d) cross-section pictures

4. Conclusions

Plasma spraying technology at atmospheric pressure offers the possibility to obtain microsized particles, granules, and coatings from inorganic metal oxides with controlled characteristics for special application. Plasma jet.particle interaction lasts for about 1.2 ms and strongly depends on jet temperature, velocity, and particle's mass.
While moving in a jet, the ceramic particle is heated, melted, and splats on the substratum. The most intense melting of particles occurs at x/d = 3.8 from exhaust nozzle.
Velocity of the particle near the substrate exceeds average plasma jet velocity and depending on the diameter of particle reaches up to 150 - 320 ms^{-1}. At x/d = 8 - 12 from exhaust nozzle the dispersed particles' flow is steady, whereas the value of kinetic energy is ultimate.
The numerical calculation data shows that the applied numerical model of two-phase high temperature jet calculation is in good agreement with experimental data and could be used to determine the optimal plasma spray parameters for coatings with desirable characteristics. The grain size of plasma sprayed coatings is smaller than 5 μm.

5. Acknowledgement

The research has been partly supported by the European Union (European Regional development Fund).

6. References

Brinkiene K. and Kezelis R. (2005) Effect of alumina addition on the microstructure of plasma sprayed YSZ*, J. Eur. Ceram. Soc. 25*, 2181-2184.

Delluc G.; Ageorges H., Pateyron B., and Fauchais P. (2005). Fast modelling of plasma jet and particle behaviours in spray conditions, *High Temp. Mater. Processes 9*, 211-226.

Delluc G.; Mariaux G.; Vardelle A.; Fauchais P. and Pateyron B. (2003). A numerical tool for plasma spraying. Part I: Modeling of plasma jet and particle behavior, in: *Abstracts and full paper CD of the ISPC 16*, Taormina, Italy, June 22-27, 2003, 6 p.

Fouchais P.; Montavon G.; Vardelle M., and Cedelle J. (2006). Developments in direct current plasma spraying, *Surf. Coatings Technol.* 201, 1908-1921 (2006).

Garbero M.; Vanni M.; and Fritsching U. (2006). Gas / surface heat transfer in spray deposition processes, *Int. J. Heat Fluid Flow 27*, 105-122.

Kersten H.; Deutsch H.; Steffen H.; Kroesen G.M.W. and Hippler M. (2001)., The energy balance at substrate surfaces during plasma processes, *Vacuum 63*, 385-431.

Kezelis R.; Valatkevicius P. and Ambrazevicius A. (1996). Velocity and temperature distribution in the entrance region of tube with high temperature turbulent air flow, *Trudy Akademii Nauk Litovskoy SSR B 6*(97), 57-61 (1976) [in Russian].

Klocker T.; Dorfmann M. and Clyne T. W. (2001). Process modelling to optimise the structure of hollow zirconia particles for use in plasma sprayed thermal barrier coatings, in: *ITSC 2001*, eds. C.C. Berndt, K.A. Khor, and E.F. Lugscheider (ASM, Singapore, 2001) 149-155.

Landes K. (2006). Diagnostics in plasma spraying techniques, *Surf. Coatings Technol.* 201, 1948-1954.

T&TWinner can be download from http://ttwinner.free.fr .

Valatkevicius P.; Kru.inskaite V.; Valinciute V. and Valincius V. (2003). Preparation of catalytic coatings for heterogeneous catalysts employing atmospheric pressure non-equilibrium plasma, *Surf. Coatings Technol. 174. 175*, 1106-1110.

Valincius V.; Krusinskaite V.; Valatkevicius P.; Valinciute V., and Marcinauskas L. (2004). Electric and thermal characteristics of the linear, sectional DC plasma generator, *Plasma Sources Sci. Technol. 13*, 199-206.

Valincius V.; Valatkevicius P. and Marcinauskas L. (2003). Preparation of hard coatings employing nonequilibrium plasma under atmospheric and reduced pressure, in: *16th International Symposium on Plasma Chemistry ISPC-16: Proceedings*, Taormina, Italy, June 22.27, 2003 (University of Bari, Italy, 2003) 1-6.

Valinciute V. (2007) (*Research on Plasma Spray Pyrolysis in the Processes of Coatings Synthesis*, Summary of the doctoral dissertation (Kaunas University of Technology, 2007).

Yeh F. B. (2006). The effect of plasma characteristics on the melting time at the front surface of a _lm on a substrate: An exact solution, *Int. J. Heat Mass Transfer 49*, 297-306.

Part 4

Electrohydraulic Systems

19

Numerical Simulation - a Design Tool for Electro Hydraulic Servo Systems

Popescu T.C.[1], Vasiliu D.[2] and Vasiliu N.[2]
[1]*National Institute for Optoelectronics, INOE 2000-IHP Bucharest,*
[2]*University „Politehnica" of Bucharest*
Romania

1. Introduction

Electro hydraulic servo systems are complex technical entities that involve both phenomena of fluid mechanics, and phenomena specific to control processes with feedback. Due to the complexity of these interactions, the optimal design goal is achieved by an iterative process, using some dedicated software. To obtain the required performance the use of mathematical modeling and numerical simulation of these systems is always very effective. In any optimal synthesis process of an electro hydraulic control systems, the analysis of the stablity is an important stage. Several methods are used to provide a good stability for such type of systems: the increase of the dead band of the control valves, the use of some additional feedback, the decrease of the flow gain of the control valve around the hydraulic null point etc.
Numerical simulation of the dynamic systems allows gathering of necessary information about their behaviour based on a mathematical models that describe those systems. Obtaining of mathematical models as close as posible to the physical phenomena that are to be reproduced or impoved is helpful in making decisions for optimization. The most recent tendences in this field regard novel concepts, such as co-simulation and real time simulation.
This chapter presents two different examples of developing a numerical co-simulation environment, based on two software packages: AMESim (LMS IMAGINE SA, 2009) and LabVIEW (National Instruments, 2009). The most important parameters investigated are the following:

a. **the influence of the variable area gradient** of an electrohydraulic flow amplifier on the stability reserves of a electro hydraulic servomecahnism (a. Ion Guta et al., 2010);
b. **a hybrid solution of modeling / simulation** of a hydrostatic transmission with mixed control (b. Ion Guta et al., 2010).

By means of AMESim software a model of an electrohydraulic servomechanism was developed, while analysis of data collected as a result of simulations in AMESim was performed by means of virtual instrumentation, using LabVIEW software. The real time use of these two simulation / programming environments can lead to the development of advanced modeling / simulation networks of complex fluid systems controlled by digital hardware, useful for optimal system design.

2. The stability of electrohydraulic servomechanisms developed with electrohydraulic amplifiers of variable area gradient

2.1 Mathematical modeling of electrohydraulic servomechanism

Mathematical model of an electrohydraulic servomechanism with position response comprises the following equations (Vasiliu & Vasiliu, 2005): Equation of slide valve displacement; Equation of position transducer; Equation of electronic comparator; Continuity equation of subsystem directional control valve-hydraulic cylinder; Equation of current generator of proportional compensator; Motion equation of hydraulic cylinder's piston; Characteristic of directional control valve with variable area gradient.

The power stage of this valve is represented by an adjustment directional control valve, with 4 ways and 3 positions, with closed critical center. For a directional control valve with variable area gradient, fig.1, we hold:

Geometrical characteristics of unit sleeve – slide valve: D=2R – diameter of main slide valve; d = 2r - diameter of circular distribution window; a – width of rectangular distribution windows; b – lenght of rectangular distribution windows.

The following notations are introduced:

$$\begin{aligned} c_{x1} &= b \\ c_{y1} &= R \cdot \arcsin\left(0.5 \cdot a / R\right) \\ c_{x2} &= b + \sqrt{r^2 - \left(c_{y1} \cdot r / c_{y2}\right)^2} \\ c_{y2} &= R \cdot \arcsin\left(r / R\right) \end{aligned} \qquad f(x) = \begin{cases} c_{y1}, & x \in (0, b) \\ \dfrac{c_{y2}}{r} \cdot \sqrt{r^2 - x^2}, & x \in (b, c_{x2}) \end{cases} \tag{1}$$

Law of variation of drainage area, depending on stroke of slide valve, is:

$$A(x) = \begin{cases} 2 \cdot c_{y1} \cdot x, & x \in (0, b) \\ 2 \cdot \dfrac{c_{y2}}{r} \cdot \int_b^{c_{x2}} \sqrt{r^2 - x^2}\, dx, & x \in (b, c_{x2}) \end{cases} \tag{2}$$

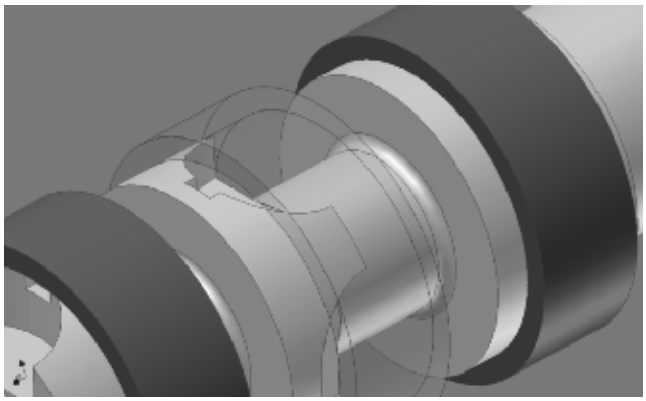


Fig. 1. Slide valve of directional control valve with variable area gradient

Slide valve of directional control valve with variable area gradient (Bosch Rexroth Group) is shown in fig.1, while variation of area of directional control valve's holes – in fig. 2. Zone I, fig.

2, is the area where drainage takes place through rectangular windows (around null), while zone II corresponds to drainage through the two distribution windows, respectively with rectangular area and quasi-elliptical area, resulted from intersection of two cylindrical bodies.

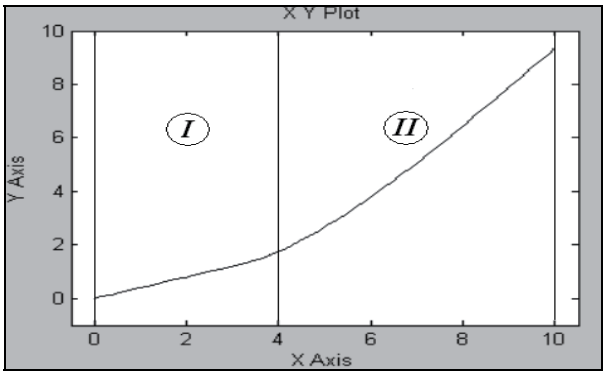


Fig. 2. Variation of area of holes, depending on relative displacement between slide valve and sleeve

Characteristics of stationary mode of directional control valve is:

$$Q = c_d \cdot A(x) \cdot \sqrt{\frac{p_s}{\rho}\left(1 - \frac{x}{|x|} \cdot \frac{p}{p_s}\right)} \tag{3}$$

For the two operation zones, flow can be calculated with the following relations:

$$Q(x) = \begin{cases} c_d \cdot 2 \cdot c_{y1} \cdot x \cdot \sqrt{\dfrac{p_s - p}{\rho}}, & x \in (0, b) \\ c_d \cdot 2 \cdot \left(c_{y1} b + \dfrac{c_{y2}}{r} \cdot \int_b^{c_{x2}} \sqrt{r^2 - x^2}\, dx \right) \cdot \sqrt{\dfrac{p_s - p}{\rho}}, & x \in (b, c_{x2}) \end{cases} \tag{4}$$

Flow amplification factor depends on operation zone:

$$K_{Qx} = \frac{\partial Q}{\partial x} = \begin{cases} c_d \cdot 2 \cdot c_{y1} \cdot \sqrt{\dfrac{p_s - p}{\rho}}, & x \in (0, b) \\ c_d \cdot 2 \cdot \dfrac{c_{y2}}{r} \sqrt{r^2 - x^2} \cdot \sqrt{\dfrac{p_s - p}{\rho}}, & x \in (b, c_{x2}) \end{cases} \tag{5}$$

Coefficient flow-pressure can be calculated with the following relations:

$$K_{Qp} = \frac{\partial Q}{\partial P} = \begin{cases} \dfrac{c_d \cdot 2 \cdot c_{y1} \cdot x}{\sqrt{\rho (p_s - p)}}, & x \in (0, b) \\ \dfrac{c_d \cdot A(x)}{\sqrt{\rho (p_s - p)}}, & x \in (b, c_{x2}) \end{cases} \tag{6}$$

$$K_p = K_{Qp} + K_1$$

Equation of slide valve displacement:
Flow control valve can be considered a delay factor of first rank:

$$\frac{x(s)}{i(s)} = \frac{K_{xi}}{T_s s + 1} \tag{7}$$

Or

$$T_s s x(s) + x(s) = K_{xi} i(s) \tag{8}$$

The following differential equation results:

$$T_s \frac{dx}{dt} + x = K_{xi} i(t) \tag{9}$$

T_s – time constant of directional control valve.
Equation of position transducer:

$$U_T = K_T y \tag{10}$$

K_T- is the constant of transducer, [V/m]
y – displacement of piston of hydraulic cylinder
Equation of electronic comparator:

$$\varepsilon = U_0 - U_T \tag{11}$$

ε - adjustment error.
Equation of current generator of proportional compensator:

$$i = K_{ie} \varepsilon \tag{12}$$

K_{ie} [A/V] –conversion factor
Continuity equation of subsystem directional control valve-hydraulic cylinder:

$$Q = A_p \dot{y} + K_l P + \frac{A_p^2}{R_h} \dot{P} \tag{13}$$

A_p -piston area;
K_l - coefficient of drainage between hydraulic cylinder's chambers;
R_h - hydraulic rigidity of double-effect hydraulic cylinder
Motion equation of hydraulic cylinder's piston
Pressure force F_p must overcome elastic force F_e, dissipation factor (the dumper) F_a, friction force, F_f and inertial force, so:

$$m_c \ddot{y} = F_p - F_a - F_e - F_f \tag{14}$$

Where,

$$F_p = A_p P \tag{15}$$

$$F_a = K_f \cdot v \tag{16}$$

$$F_e = 2\left(K_{e1} + K_{e2}\right)\left(y + y_{0e}\right) = 2K_e\left(y + y_{0e}\right) \tag{17}$$

For friction force between piston and cylinder a static component F_{fs} and a viscous one F_{fv} are both considered:

$$F_{fs} = F_{fs0} sign\dot{y} \tag{18}$$

$$F_{fv} = K_{fv}\dot{y} \tag{19}$$

2.2 Numerical co-simulation. Identification of a linearized model

Identification aims at determining static and dynamic characteristics of processes. By identification it is understood the procedure of determining a system based on one input and one output, in case of SISO systems (single input - single output), so that it is equivalent, in a certain way, to the tested system.

Identification of parameters of mathematical model based on experimental data implies four stages: acquisition of input/output data; choosing structure of the model; estimation of parameters of the model; validation of the identified model (validation of structure and value of parameters) (Calinoiu et al., 1998).

For the analyzed case we used identification procedures of ARX models (functions which use the method of least squares) in LabVIEW. ARX models have the following structure:

$$A(q)y(t)=B(q)u(t-nk)+e(t) \tag{20}$$

The identified models were the basic elements of the study, with their support Bode diagrams and transfer loci of the analyzed process are drawn.

Study of stability of automatic electro hydraulic systems can be performed based on algebraic criterion Routh - Hurwitz, which has only one condition for stability or on Nyquist criterion, which allows in addition analysis of stability reserves (Catana et al., 1996). Transfer locus of open circuit system looks like in fig. 3. Necessary and sufficient condition for closed circuit system to be stable is that the hodograph of linear model not surround the critical point (-1, j0) in the complex plane when the frequency belongs to the interval (0, ∞). Stability reserve of the system can be evaluated by two characteristic sizes: the amplitude edge (stability reserve in modulus) and the phase edge (stability reserve in phase).

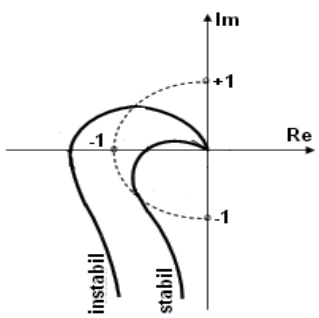


Fig. 3. Transfer locus of servo mechanism

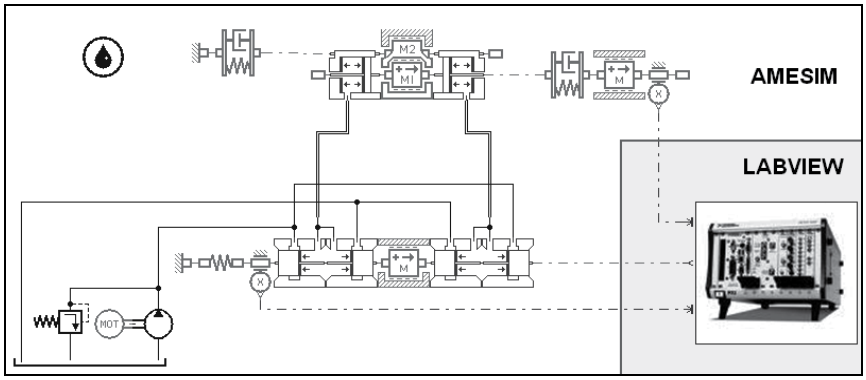


Fig. 4. Co-simulation network of analyzed servo mechanism

In fig.4 can be noticed the co-simulation network of the analyzed system. Numerical model developed in the AMESim allows analysis of behaviour over time of the examined system. Based on response to various excitation signals can be identified, by means of the model developed in LabVIEW, mathematical linearized model of the system, based on which can be performed system stability analysis.

To study system stability the transfer locus of servo mechanism was used by means of Nyquist outline analysis. The exchange of information between submodel of dynamic system of servo mechanism, developed in AMESim and compensator submodel, implemented in LabVIEW, can be achieved by shared access to a specific memory area if the networks run on the same system or by a communication channel TCP / IP if the networks run on two different systems.

Architecture of the process is *master / slave* type, the integration step is determined by the master system.

Co-simulation network, fig.4, is made of: the group of oil supply under constant pressure (constant speed electric motor, volumetric pump, normally closed valve); electro-hydraulic directional control valve with variable area gradient; linear hydraulic motor with double effect and double rod; inertial load; displacement transducer for slide valve of distributor; displacement transducer for the hydraulic cylinder rod; control software interface, analysis and interpretation of data, developed in LabVIEW.

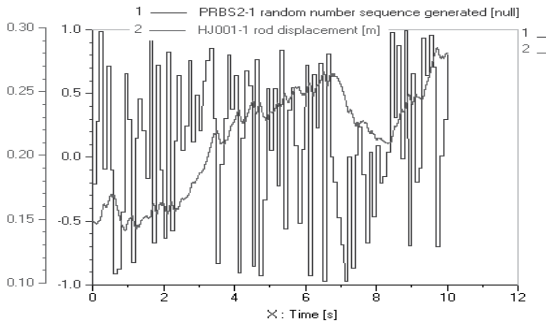
The calculations were performed for these data: m_load = 100 Kg, p_supply = 160 bar, cylinder_stroke = 300 mm, d_cylinder = 26 mm, d_rod = 12 mm, anchor rigididity = 2.1 · 107 N/m, damping coefficient = 4000 Ns/m.

The model was excited with signals type *white noise*. To establish the transfer locus, the model was examined in open loop, after drawing features we also developed chart of response over time to step closed-loop signals. ARX models identified were determined for each operating mode. Results of co-simulation are presented in figures 5, 6, 7, 8 and 9.

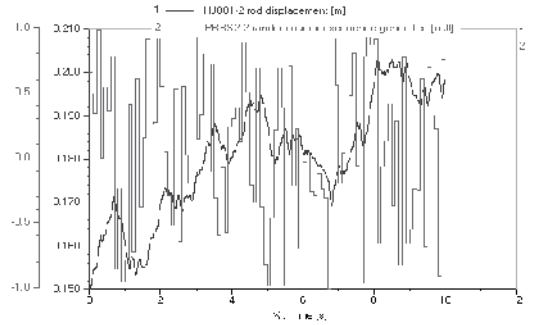
Discret mathematical models identified for open-loop system with: (a) – directional control valve without variable area gradient, (b)- directional control valve with variable area gradient:

$$(1 - 2.9719z^{-1} + 2.9479z^{-2} - 0.976z^{-3})\, y(k) = (-2.3882E\text{-}6 + 1.6886E\text{-}6z^{-1} + 3.2132E\text{-}5z^{-2})\, u(k) + e(k) \quad (a)$$

$$(1 - 2.981z^{-1} + 2.9668z^{-2} - 0.9858z^{-3})\, y(k) = (-6.9855E\text{-}7 + 5.3387E\text{-}7z^{-1} + 1.8398E\text{-}5z^{-2})\, u(k) + e(k) \quad (b)$$

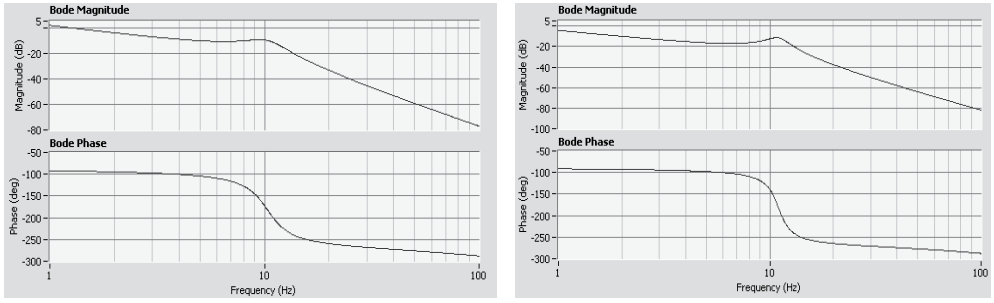


a) with constant area gradient



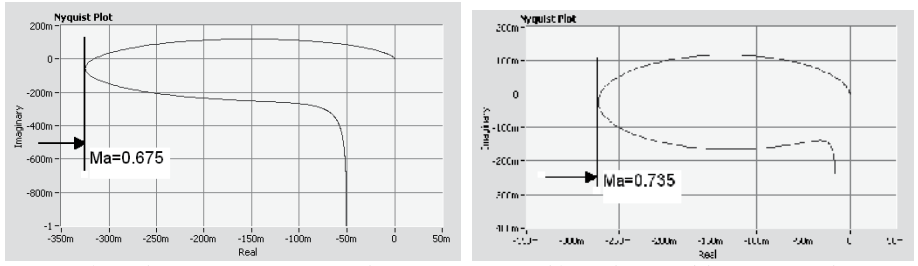
b) with variable area gradient

Fig. 5. Response over time of the servo mechanism to control signal of type white noise



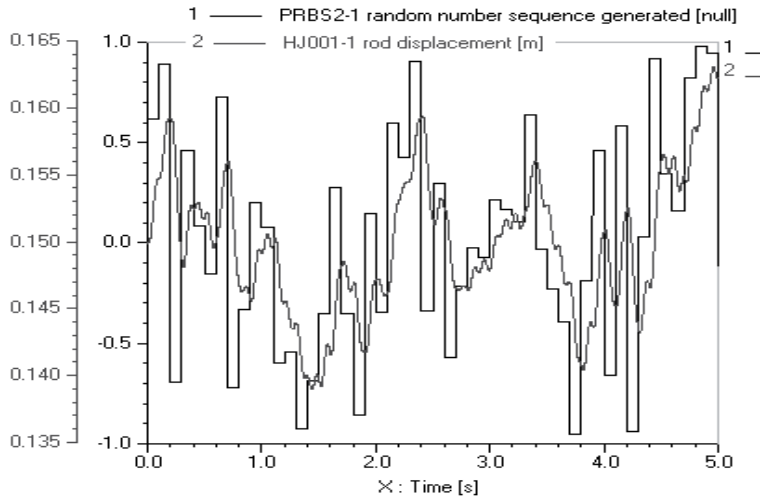
a) with constant area gradient b) with variable area gradient

Fig. 6. Bode diagram of the servo mechanism in open loop

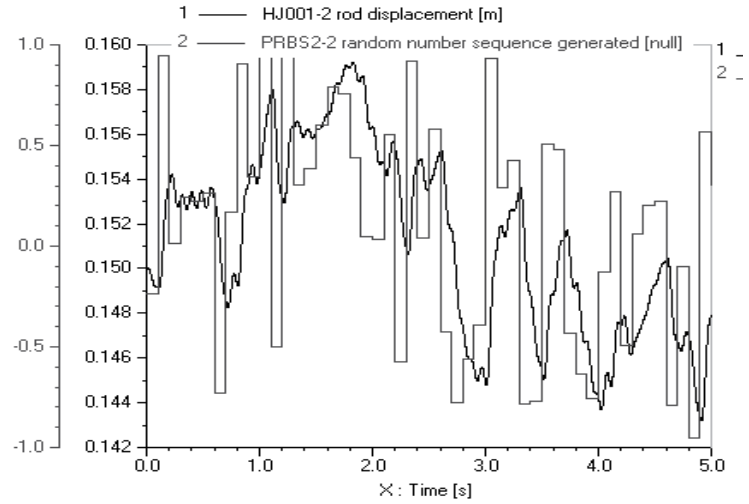


a) with constant area gradient b) with variable area gradient

Fig. 7. Hodograph of open-loop system

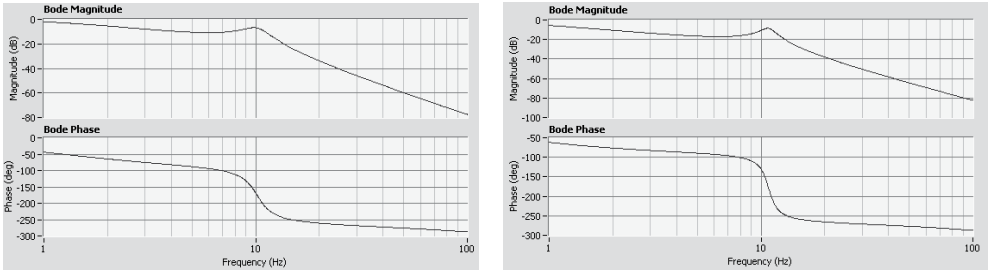


a) with constant area gradient



b) with variable area gradient

Fig. 8. Response over time of system in loop of response to control signal of type white noise



a) with constant area gradient b) with variable area gradient

Fig. 9. Bode diagram of the servo mechanism in closed loop

Discret mathematical models identified for closed-loop system with: (a) – directional control valve without variable area gradient, (b)- directional control valve with variable area gradient:

$$(1 - 2.9727z^{-1} + 2.9495z^{-2} - 0.9768z^{-3})\, y(k) = (-1.8146E\text{-}6 + 1.2474E\text{-}6z^{-1} + 3.0294E\text{-}5z^{-2})\, u(k) + e(k) \quad (a)$$

$$(1 - 2.9819z^{-1} + 2.9685z^{-2} - 0.9866z^{-3})\, y(k) = (-5.7764E\text{-}7 + 5.2905E\text{-}7z^{-1} + 1.742E\text{-}5z^{-2})\, u(k) + e(k) \quad (b)$$

3. Optimization of hydrostatic transmissions by means of virtual instrumentation technique

3.1 Problem formulation

The analyzed hydrostatic transmission, of mixed adjustment, with single consumer of type adjustable rotary volumetric motor, according to the basic model in fig.10, includes:

- **in its primary sector**: a MOOG servopump, place 1, with radial pistons and integrated electronics, with three loops of adjustment, that is in flow, in pressure, in flow and pressure, with capacity of 32 cm^3/rev, rotary speed of 1450 rev/min, control voltage of 0...10V, flow of 0...46 l/min; an electric motor for servopump actuation, of constant rotary speed, place 2; a pressure limiting valve, place 3; a flow transducer, place 4; and a pressure transducer, place 5.
- **in its secodary sector**: a BOSCH servo motor type EP2, place 7, with axial pistons, tilted block and integrated electronics, with minimum capacity of 7 cm^3/rev at control voltage of 200 mA and maximum capacity of 28 cm^3/rev at control voltage of 800 mA, at supply voltage of 24Vd.c.; a torque transducer place 8; a speed transducer, place 9; an axial piston pump, with tilted block and fixed capacity place 10, to simulate the load of hydraulic servo motor; two pressure transducers, place 11 and place 13; four way-valves, place 12, fitted on suction / repression side of load pump; a pressure adjustment valve, with electric control, place 14, for adjusting load of hydraulic servo motor.
- **a** PXI-NATIONAL INSTRUMENTS **block**, place 6, which provides a virtual interface of the adjustment process of capacity of the adjustable volumetric machines (LabVIEW / PXI).

  For this hydrostatic transmission we have developed a physical laboratory model, according to fig.11 and fig.12; a numerical simulation network in AMESim, according to fig.13; a virtual interface for the adjustment model of transmission, according to fig.14 and a web interface for the adjustment model of transmission, according to fig.15.

By means of the adjustment model of hydrostatic transmission with mixed adjustment the following items were highlighted:

- **by means of co-simulation AMESim-LabVIEW**: demonstration, on the physical laboratory model, of primary (pump), secondary (motor) and mixed (pump and motor) adjustments, specific to hydrostatic transmissions (Popescu et al., 2010); demonstration, on the physical laboratory model, of the advantages, in terms of energy, of hydrostatic transmissions with adjustable pumps in their primary sector over those with fixed pumps in their primary sector (b. Drumea et al., 2010).
- **by means of simulation models in AMESim**: optimization of the adjustment model of a hydrostatic transmission with mixed adjustment in order to reduce variation of the rotary speed of volumetric motor within its secondary sector, in accordance with variation of its load.

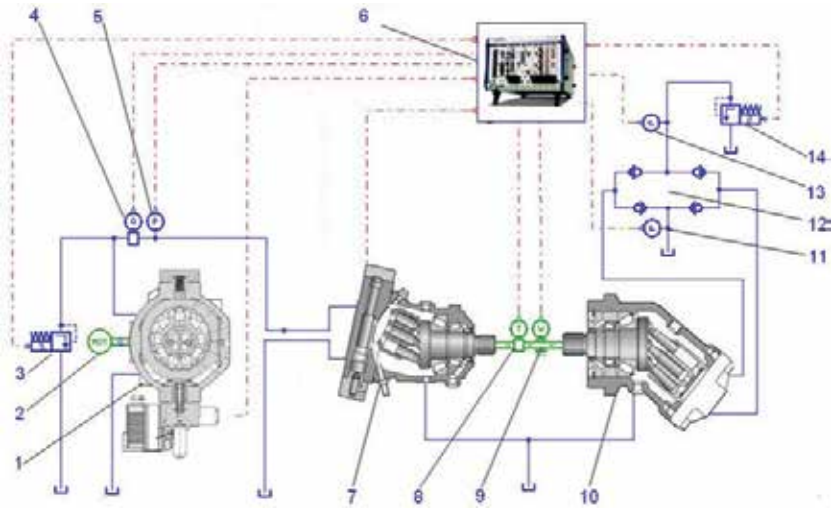


Fig. 10. Basic model of a hydrostatic transmission with mixed adjustment

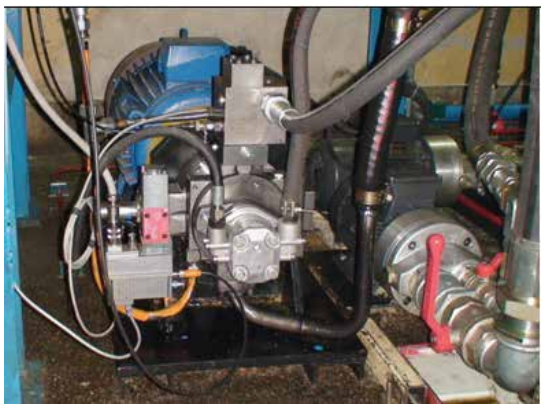


Fig. 11. MOOG servopump, type RKP-D, within the primary sector

Fig. 12. Bosch servo motor, type EP2, within the secondary sector

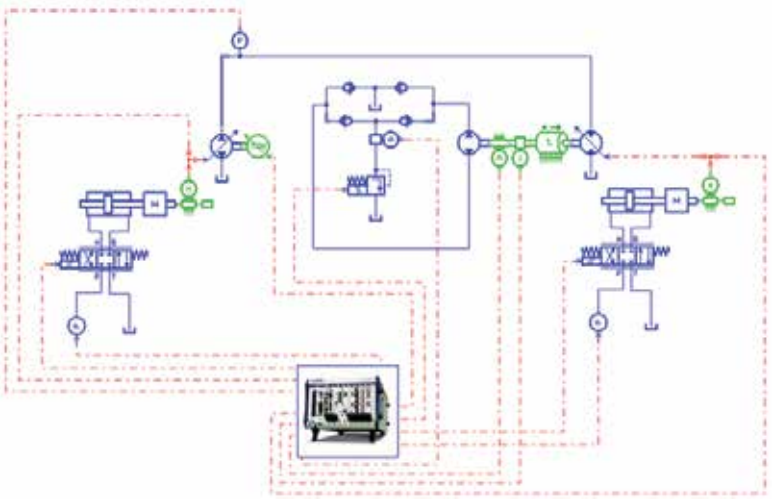


Fig. 13. Simulation network in AMESim of a hydrostatic transmission with mixed adjustment

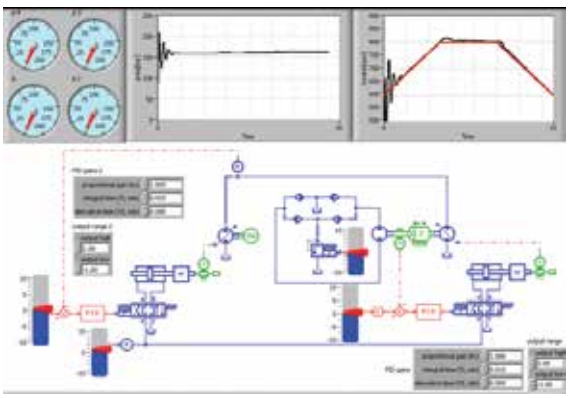


Fig. 14. Virtual interface of the adjustment model (LabVIEW / PXI)

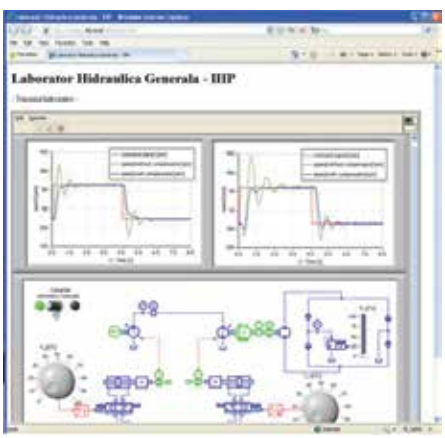


Fig. 15. Web interface of the adjustment model (LabVIEW / PXI)

3.2 Demonstration of adjustment of capacity at adjustable volumetric machines

We have traced the response of the adjustment system of transmission, that actuates upon the servomechanism which adjusts the capacity of the pump within the primary sector or upon the servomechanism which adjusts the capacity of the motor within the secondary sector, to rotational speed step type signal imposed to the volumetric motor within the secondary sector. Within the adjustment model we have preset the rotational speed threshold of 320 r.p.m, below which the adjustment of transmission is performed upon the pump (primary adjustment) and above which the adjustment of transmission is performed upon the motor (secondary adjustment).

Dynamic characteristics of the system were raised, which highlight:

- the influence of a rotational speed step type signal of 312 rpm imposed to the hydraulic motor within the secondary sector, upon the adjustment drive of capacity of the adjustable pump, with and without erorr compensation, according to fig.16;
- the influence of a rotational speed step type signal of 410 rpm imposed to the hydraulic motor within the secondary sector, upon the adjustment drive of capacity of the adjustable pump, with and without erorr compensation, according to fig.17;
- the influence of rotational speed step type mixed signals of 308 rpm, respectively 408 rpm, imposed to the adjustable hydraulic motor within the secondary sector, upon the adjustment drive of capacity of the adjustable pump, respectively of the adjustable motor, with and without erorr compensation, according to fig.18.

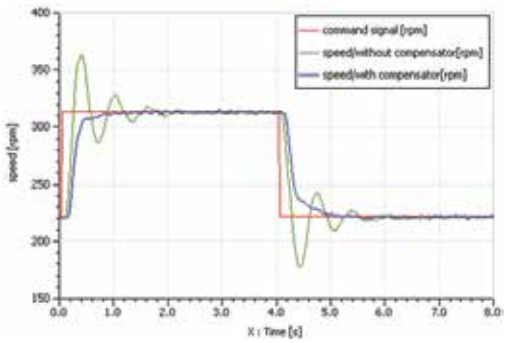


Fig. 16. Response of the adjustment system of rotational speed of hydraulic motor to step type excitation signal – pump capacity drive

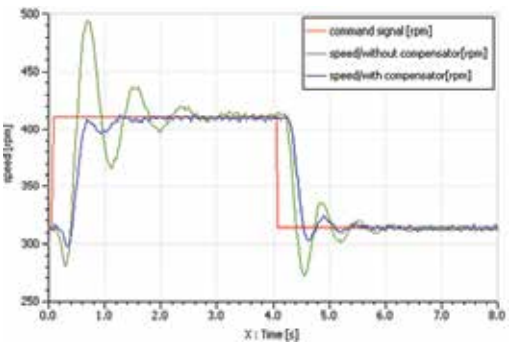


Fig. 17. Response of the adjustment system of rotational speed of hydraulic motor to step type excitation signal – motor capacity drive

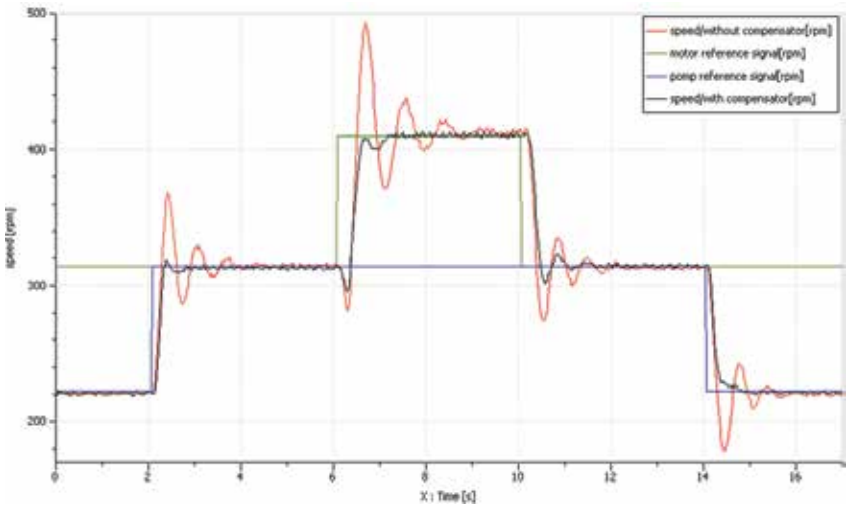


Fig. 18. Response of the adjustment system of rotational speed of hydraulic motor to mixed step type excitation signal – pump and motor capacity drive

3.3 Numerical simulation of radial piston pump MOOG type RKP-D

Simulation model developed for the analysis of volumetric pump in fig.19, is shown in fig.20 (a. Drumea et al., 2010). It includes: the hydraulic servomechanism for prescribing the position of the adjustment ring; a module for calculating the relative position of small pistons as against to their angular position and the ordered eccentricity; the two small radial pistons of the pump; the distribution unit, controlled by the angular position of small pistons and the geometrical characteristics of the distribution flange.

By means of the modeling network developed, dynamic characteristics of the servo motor that adjusts capacity of the analyzed radial piston pump were determined, figures 5 and 6. The model was excited with control signals (prescribing of eccentricity of the flow adjustment / control ring), triangular, sinusoidal and rectangular signals, of various amplitudes and frequencies. Obtained results are compared, simulated and experimentally shown. Simulation model has been "tuned" as a result of the comparative analysis between simulated and experimental response for a better accuracy of results (Popescu et al., 2010).

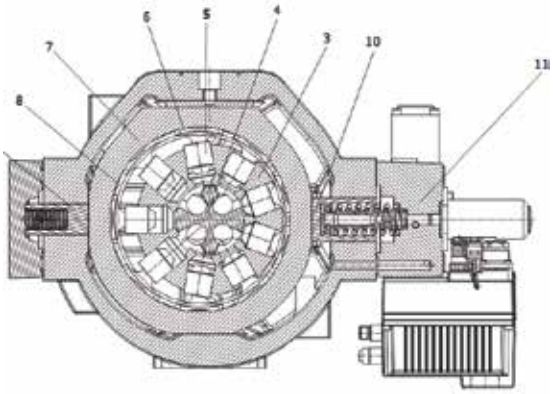


Fig. 19. Servo pump MOOG type RKP-D; cross-section

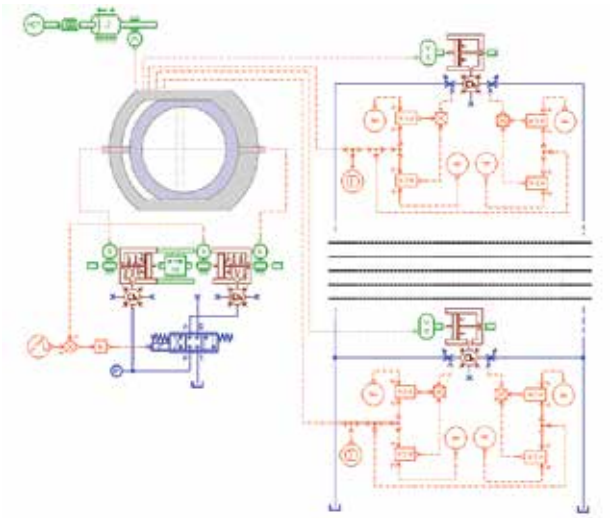
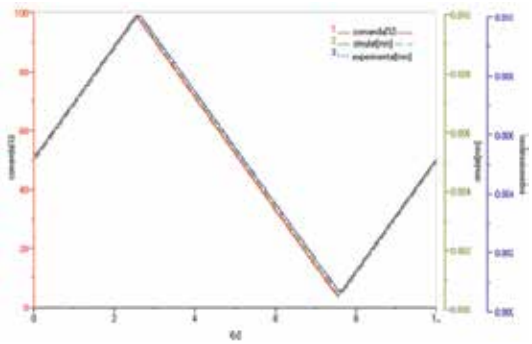
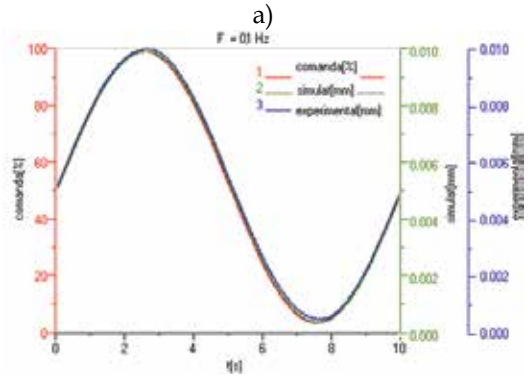


Fig. 20. Servo pump MOOG type RKP –D; numerical simulation model

In figures 21 and 22 curve 1 represents the control signal, curve 2 - response of servomechanism that adjusts capacity, obtained through numerical simulation, and curve 3 - response of servomechanism that adjusts capacity, obtained on an experimental basis.



a)



b)

Fig. 21. Response of the adjustment servomechanism to control triangular, (a) and sinusoidal, (b) signals (f=0.1 Hz)

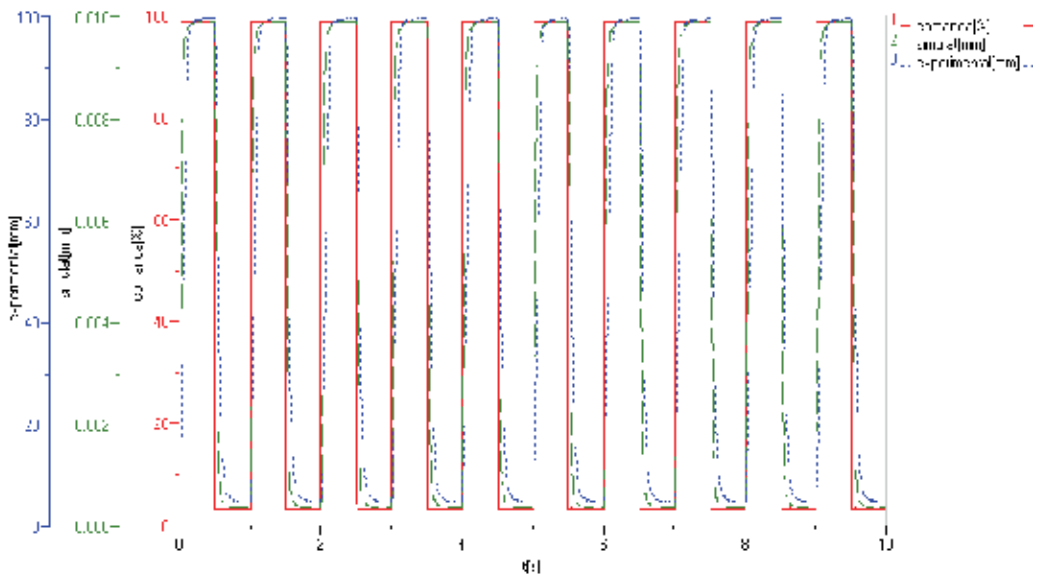


Fig. 22. Response of the adjustment servomechanism to a control rectangular signal (f=1 Hz)

3.4 Demonstration of advantages in terms of energy of hydrostatic transmissions with adjustable pumps in their primary sector

Two variants of adjusting the flow within the primary sector of a hydrostatic transmission were tested in comparison, according to fig.23.

- variant (a), where capacity of the pump within the primary sector is set, while flow adjustment is performed by means of an adjustable throttle (Popescu et al., 2009). In this case, the extra flow is discharged at the tank through a normally closed pressure valve;
- variant (b), where in the primary sector a hydraulic servopump with adjustable capacity is used.

In both variants, tests were performed for a constant load of 20 Nm at the shaft of hydraulic motor within the secondary sector of transmission.

After calibrating the adjustment model of mixed adjustment transmission, process carried out by means of the numerical simulation network, the motor within the secondary sector was set on maximum capacity and the two flow adjustments systems for the pump in the primary sector were analyzed comparatively (b. Drumea et al., 2010). Tests were performed for a 20 Nm load at the shaft of the motor within secondary sector of transmission.

Experimentally, systems were excited cyclically, with or without energy efficiency, with step-type control signals of rotational speed (500 rpm), fig.24, and ramp-type signals, fig.25. We recorded evolution over time of rotational speed of the hydraulic motor shaft, fig.24 (a), fig.25 (a) and pressures within primary hydraulic circuit, fig.24 (b), fig.25 (b).

The two systems for adjusting the flow within the primary sector of transmission have been excited with a control signal of the speed of hydraulic motor within the secondary sector, corresponding to a specific preset profile.

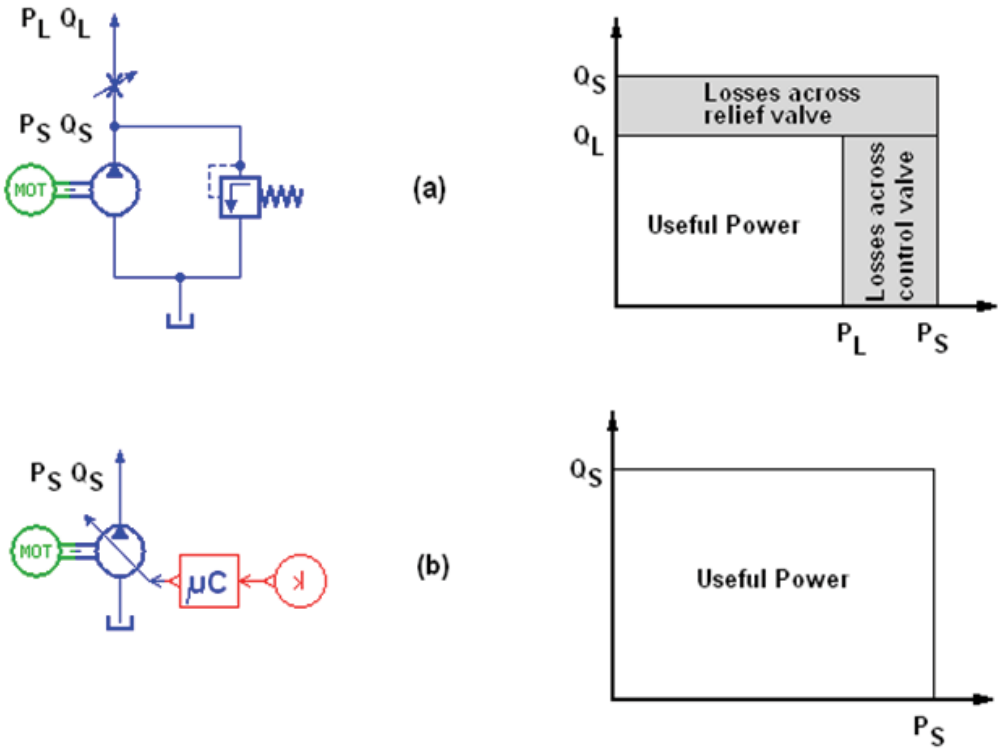
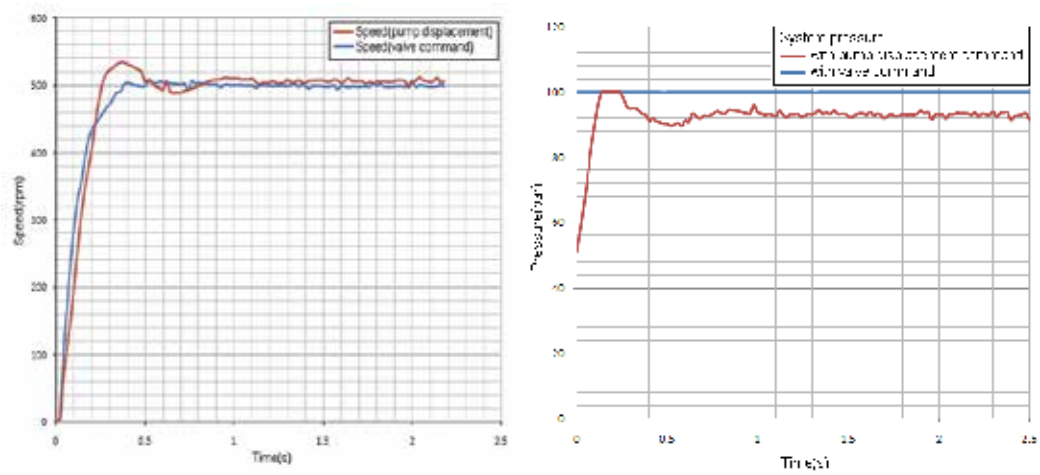
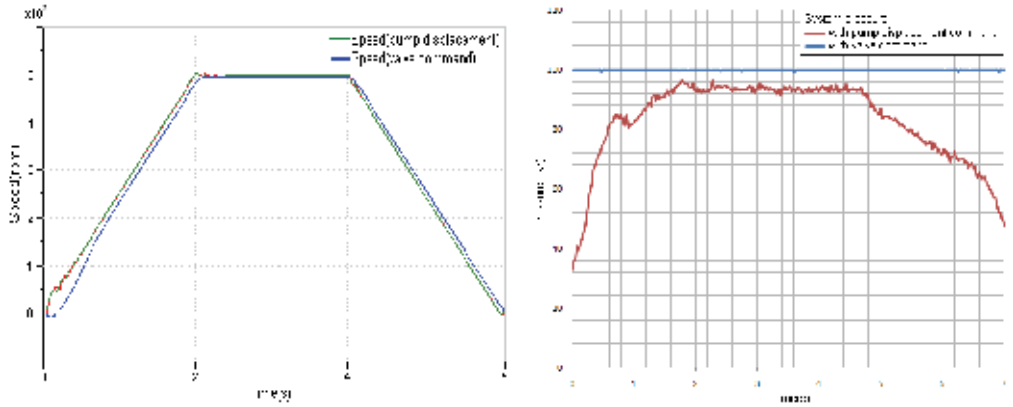


Fig. 23. Variants of adjusting the flow within the primary sector of a hydrostatic transmission



a) b)

Fig. 24. a) Variation of rotational speed of hydraulic motor within secondary sector to a step type excitation signal of flow adjustment systems within the primary sector, b) Variation of pressure along the primary circuit to a step type excitation signal of flow adjustment systems within the primary sector



a) b)

Fig. 25. a) Variation of rotational speed of hydraulic motor within the secondary sector to a ramp type excitation signal of flow adjustment systems within the primary sector, b) Variation of pressure along the primary circuit to a ramp type excitation signal of flow adjustment systems within the primary sector

We have recorded the pressures within the circuit, fig.26 (blue- variation of rotational speed of hydraulic motor; brown- variation of the supply voltage of hydraulic motor in an energy-efficient system; red- variation of the supply voltage of hydraulic motor in an energy-inefficient system).

In fig.27, after calculation of hydraulic power used by the pump within the primary sector (P=Q*p), we have traced evolution over time of this power for the two flow adjustment systems, without (brown) and with (blue) energy efficiency.

The obtained data were integrated numerically in order to result the evolution of the consumed energy, fig.28 (brown- energy-inefficient system, blue- energy-efficient system). Area of the surface defined by the two curves represents the energy savings.

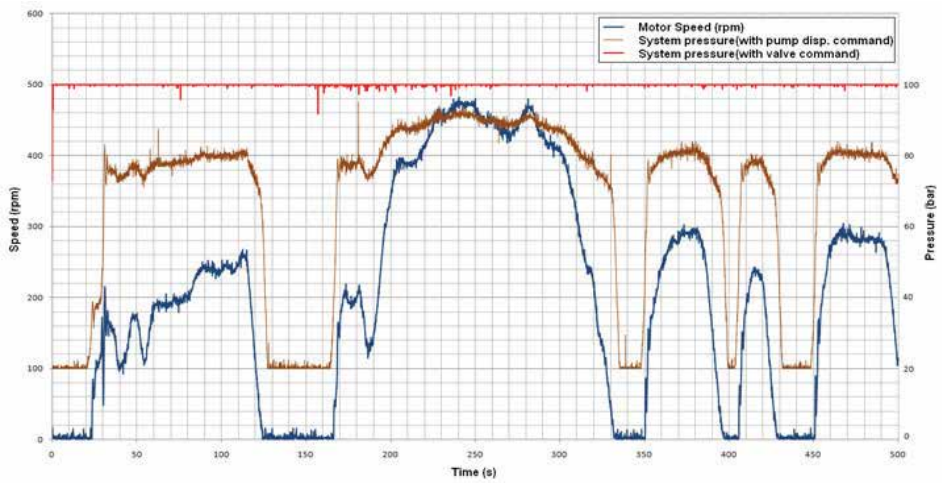


Fig. 26. Variation of pressure along the primary circuit of hydraulic transmission, at control signal with preset profile for rotational speed of the motor within the secondary sector

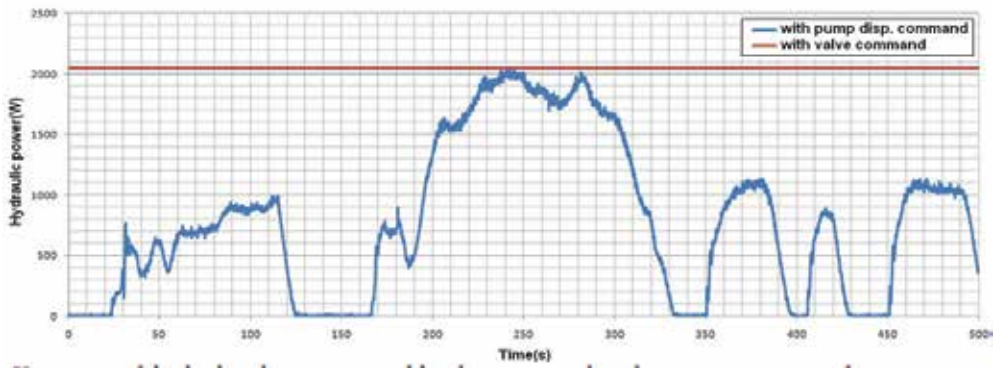


Fig. 27. Variation of the hydraulic power used by the pump within the primary sector of transmission, at control signal with preset profile for rotational speed of the motor within the secondary sector

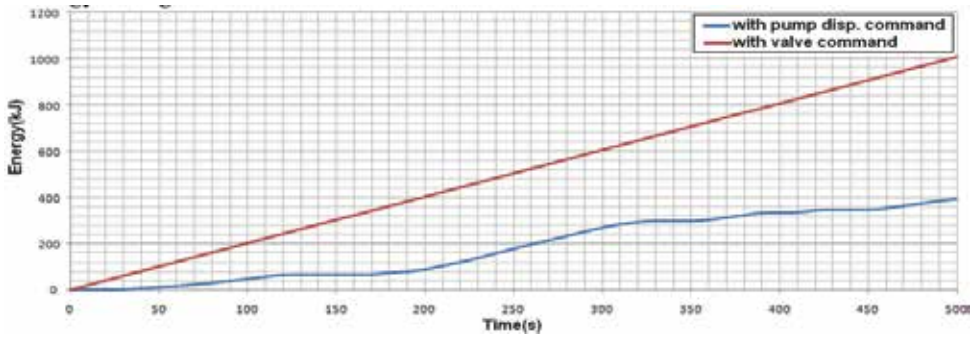


Fig. 28. Variation of energy used by the pump within the primary sector of transmission, at control signal with preset profile for rotational speed of the motor within the secondary sector

3.5 Optimization of the adjustment model of a hydraulic transmission with secondary adjustment

We aimed to optimize the adjustment model of a hydraulic transmission with secondary adjustment, derived from the hydraulic transmission with mixed adjustment. Optimization is performed in view of two goals: to reduce the variation range of rotational speed of the motor within the secondary sector, caused by variation of its load, and to reduce the extra flow discharged through the valve of the pump within the primary sector of transmission.

A pre-step in optimizing the adjustment model of hydrostatic transmission is represented by optimization of the simulation model in AMESim of transmission. In order to achive this, three variants of simulation models were developed:

- **variant (a)**, according to fig.29: hydraulic transmission with secondary adjustment, with fixed pump and variable motor, with compensator type P (proportional) in the adjustment loop of rotational speed;
- **variant (b)**, according to fig.30: hydraulic transmission with secondary adjustment, with fixed pump and variable motor, with compensator type PID (proportional, integrative, derivative) in the adjustment loop of rotational speed;

- **variant (c)**, according to fig.31: hydraulic transmission with secondary adjustment, with adjustable pump equipped with pressure regulator and variable motor, with compensator type P (proportional) in the adjustment loop of rotational speed.

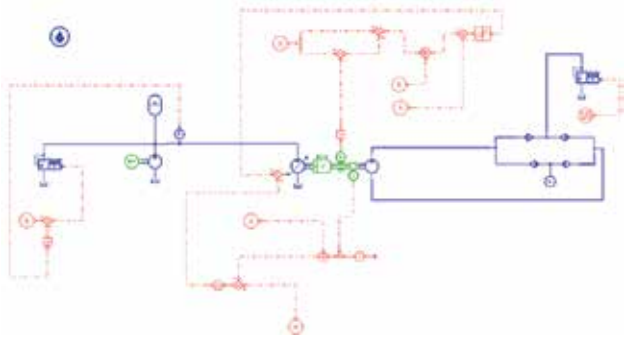


Fig. 29. Simulation model in AMESim – hydrostatic transmission with secondary adjustment: **variant (a)**

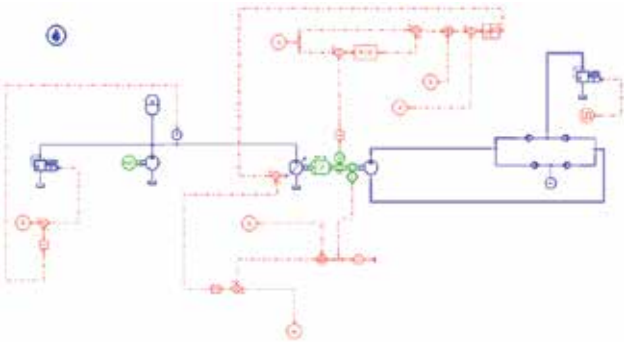


Fig. 30. Simulation model in AMESim – hydrostatic transmission with secondary adjustment: **variant (b)**

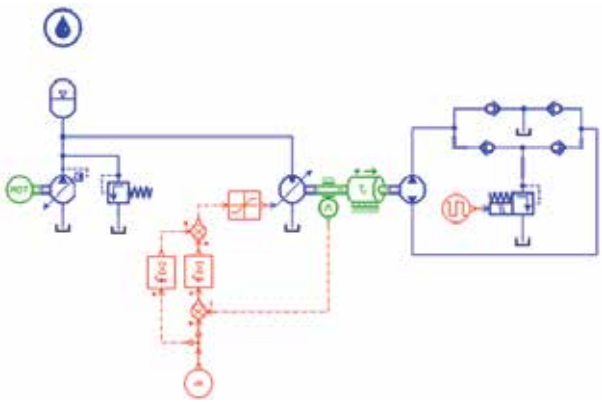


Fig. 31. Simulation model in AMESim – hydrostatic transmission with secondary adjustment: **variant (c)**

For the three simulation models the proportional pressure valve, associated with the fixed pump for simulation of the load of the servo motor witin the secondary sector of hydrostatic transmission, is excited with a rectangular signal. We traced variation over time of the rotational speed of hydraulic servo motor, caused by variation of its load. Simulation model – variant (c) is run simultaneously for three maximum values of the capacity of hydraulic servo motor. Obtained results are presented in fig.32, fig.33 and fig.34.

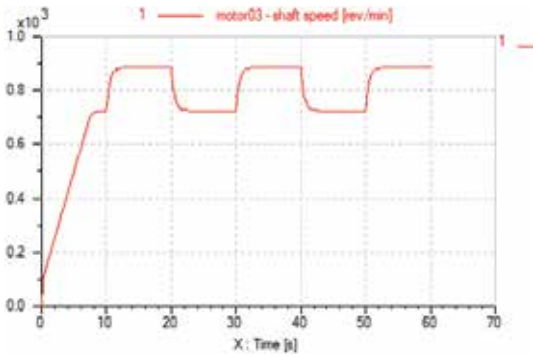


Fig. 32. Variation of rotational speed of hydraulic motor-**variant (a)**

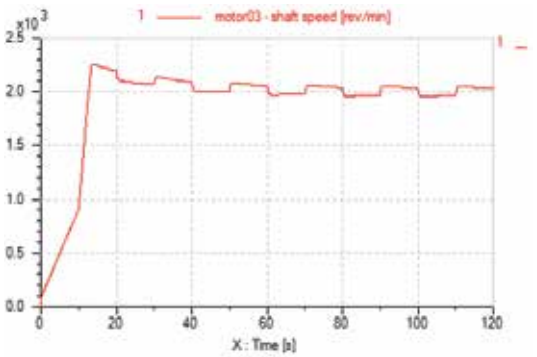


Fig. 33. Variation of rotational speed of hydraulic motor- **variant (b)**

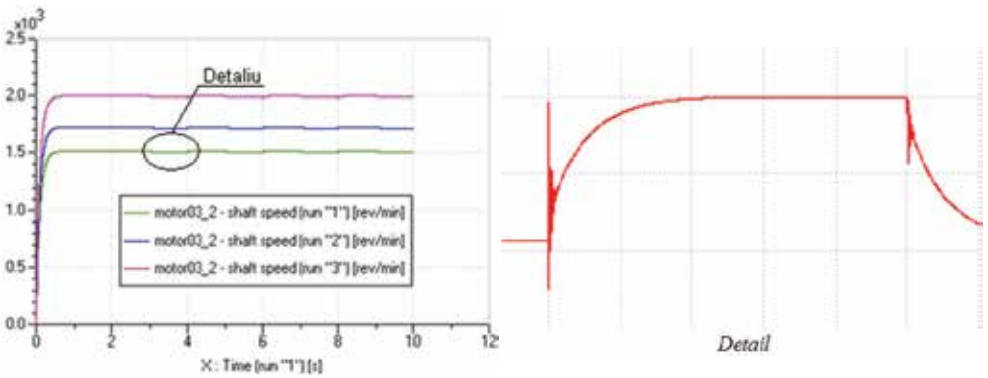


Fig. 34. Variation of rotational speed of hydraulic motor- **variant (c)**

4. Conclusions

Obtaining mathematical models, as close as possible to physical phenomena which are intended to be replicated or improved, help us in deciding how to optimize them. The introduction of computers in monitoring and controlling processes caused changes in technological systems. With support from the methods for identification of processes and from the power of numerical computing equipment, researchers and designers can shorten the period for development of applications in various fields by generating a solution as close as possible to reality, since the design stage.

4.1 Conclusions

As you can see from the diagrams of the transfer locus of servo mechanism (fig.7), its stability reserve increases along with decreasing amplification factor in flow of directional control valve around the null.
It can be noted that the amplification margin increases by about 15 percent when using directional control valve with variable area gradient, approaching the value 0.75, which is considered optimal in the literature (Vasiliu & Vasiliu, 2005).
Authors of this paper consider interesting the approach to simulate the method of enlarging the stability reserve by reducing the amplification factor around null and conditioning the control signal of distributing electromagnet to "trace" area profile of a slide valve "geometrically" shaped.
The advantages of using this method are: reduced cost of execution of slide valves, possibility to upgrade for systems without variable area gradient, improvement of performance of servo systems.

4.2 Conclusions

The adjustment model of a hydrostatic transmission, developed through technique of co-simulation AMESim/ LabVIEW, enables virtual and experimental analysis of phenomena specific to fluid power installations.
The web adjustment model of the hydrostatic transmission enables access to the drives and results of tests carried upon the physical laboratory model also for the persons outside the laboratory.
The adjustment model of a hydrostatic transmission with mixed adjustment enables remote control upon the parameters of hydraulic motor, depending on the actuation on hydraulic servomechanisms that adjust capacities of adjustable volumetric machines within the primary and secondary sectors.
The developed adjustment model highlights the advantages in terms of energy of hydraulic transmissions with servopumps within their primary sector over transmissions with fixed pumps within their primary sector.
By means of successive simulations in AMESim, on three simulation models equivalent to the physic laboratory model, a hydraulic transmission with secodary adjustment was optimized, in terms of energy consumption and functional performances.
Optimal version of the simulation model, variant (c), represents the basis for optimization of the adjustment model for hydraulic transmissions with secodary adjustment, derived from the hydraulic transmission with mixed adjustment. In this variant of simulation the variation range of rotational speed of the hydraulic motor within the secondary sector, in conditions of variation of its load after a rectangular signal, is minimum, while through the pressure valve of the pump within the primary sector the extra flow is null.

5. References

Bosch Rexroth Group - www.boschrexroth.com

Calinoiu, C., Vasiliu, N. & Vasiliu, D. (1998). Modeling, Simulation and experimental Identification of the Hydraulic Servomechanisms, *Technical Publishing House*, Bucharest, Romania, 222 p., ISBN 973-31-1315-8

Catana, I.; Vasiliu, D. & Vasiliu, N. (1996). Servomecanisme electrohidraulice, *Technical Publishing House,* POLITEHNICA University, Bucharest

a. Drumea, P.; Popescu, T.C.; Blejan, M. & Rotaru, D. (2010). Research Activities Regarding Secondary and Primary Adjustment in Fluid Power Systems, *7th International Fluid Power Conference Aachen Efficiency through Fluid Power, Scientific Poster Session,* Aachen, Germany, 22-24 March 2010

b. Drumea, P.; Popescu, T.C. & Ion Guta, D.D. (2010). Research activities regarding energetic and functional advantages of hydraulic transmissions, *Proceedings of SGEM 2010*, 10th International Multidisciplinary Scientific Geo-Conference & EXPO Modern Management of Mine Producing, Geology and Environmental Protection, Albena Resort-Bulgaria, 20 June - 25 June 2010

a. Ion Guta, D.D.; Lepadatu, I.; Popescu, T.C. & Dumitrescu, C.(2010). Research on the stability of electrohydraulic servomechanisms developed with electrohydraulic amplifiers of variable area gradient, *Proceedings of 2010 International Conference on Optimisation of the Robots and Manipulators,* Calimanesti, Romania, 28-30 May, 2010

b. Ion Guta,D.D.; Popescu,T.C. & Dumitrescu,C.(2010). Optimization of hydrostatic transmissions by means of virtual instrumentation technique, *Proceedings of ATOM-n 2010,* The 5th edition of the International Conference "Advanced Topics in Optoelectronics, Microelectronics and Nanotechnologies", 26-29 August 2010, Constanta, Romania

LMS IMAGINE SA (2009). Advanced Modelling And Simulation Environment, Release 8.2.b., *User Manual*, Roanne, France

National Instruments (2009). *User Manual,* NI USB-621x

Popescu, T.C.; Ion Guta, D.D. & Marin, A. (2010). Adjustment of hydrostatic transmissions through virtual instrumentation technique, ENERG_02, *Proceedings of ISC 2010*, June 7-9, 2010, Budapest, Hungary

Popescu, T.C.; Lepadatu,I. & Ion Guta,D.D. (2009). Experimental research activies regarding the reduction of energy compsuntion at endurance test stands of rotary voliumetric machines, *Proceedings of* International Scientific Technical Conference *Hydraulics and Pneumatics 2009*, Wroclaw, 7-9 October, ISBN 978-83-87982-34-8, pp. 303-310.

Vasiliu, N. & Vasiliu, D. (2005). Fluid Power Systems, Vol.I., *Technical Publishing House,* Bucharest, Romania, ISBN 973-31-2249-1

20

Applications of the Electrohydraulic Servomechanisms in Management of Water Resources

Popescu T. C.[1], Vasiliu D.[2], Vasiliu N.[2] and Calinoiu C.[2]
[1]*National Institute for Optoelectronics, INOE 2000-IHP Bucharest,*
[2]*University „Politehnica" of Bucharest*
Romania

1. Introduction

Water and energy are key component of almost all human activities. Water supply is vital to feeding the growing world population, to production of mat natural systems on which life on earth relies. In the context of erial goods which cause rising of living standards and to maintaining the integrity of **world water crisis**, more and more governments have begun to develop new policies for future, aiming at rationalization and efficiency of water consumption.

Maximum reduction of waste and undue loss of water is an important objective in the management of water resources. In this sense any valuable technical solution that helps achieve these goals, especially at the large water consumers, deserves to be implemented. To illustrate this, we propose the use of the automated land leveling systems, tracing a laser reference plane, as a method of reducing water losses in two activity areas: crop irrigation and construction of earth dams in hydropower stations. In the last decade, these works are performed with leveling machines equipped with laser modular systems manufactured by companies like TOPCON from Japan or APACHE and SPECTRA PRECISION from USA.

In the first kind of activity, the solution ensures equal conditions, in terms of water consumption, for all plants on agricultural land, watered by natural way - rainfalls, or artificially - by irrigation systems. This solution prevents areas of "pools" of water on agricultural land, and providing deviations from the reference plane of max. ± 2 cm by only 2 passes of the leveling machine (rough, and smoothing leveling) eliminates soil loosening works, which are necessary after conventional leveling.

In the second case, the method ensures optimal thickness, depending on the type of compression equipment, of the earth layers deposited when constructing dam body, with maximum deviations of ± 2 cm over the entire surface of the deposited layer. This type of smoothing performed before the compaction of each layer of soil deposited in the dam body, ensures a uniform degree of compaction of the dam, fig.1. Homogeneity of dam compaction is a measure for reducing seepage through the dam and pronounced settlements of the top, causing possible overflowing of water above the top of the dam, fig.2.

Laser modulation systems are not included in the standard facilities, not even in the latest modern leveling machines, fig.3. They can be installed on any hydraulically actuated

leveling machine, regardless of its wear or origin. The laser modular system includes a laser emission module, a laser receiver module, an electronic block and an electro hydraulic servo system (distribution block). Starting in 2008, the mounting on leveling equipment and the service of TOPCON laser modular equipment is performed in Europe by the representatives of Geodis Brno, Geodis Slovakia, Geodis Geodis Ro and Geodis Austria. These distributors are companies that get the laser and electronic modules from TOPCON, and the electro-hydraulic components from different suppliers.

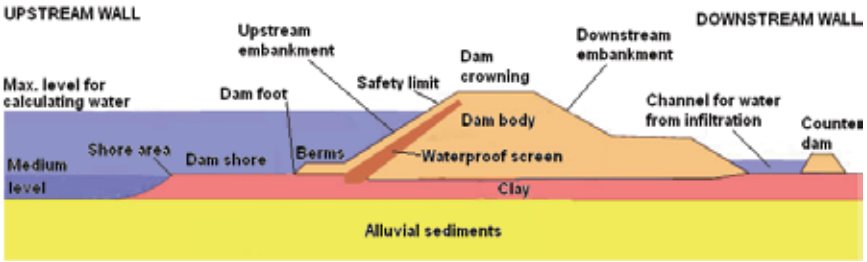


Fig. 1. Main section through an earth dam

Fig. 2. Example of water flood over a dam January the 1st 2006 Sherman Island

Fig. 3. Laser controlled electro hydraulic leveling machine (courtesy New Holland)

The wide diversity of the fluid power systems used on leveling machinery, supplied by various manufacturers, and with different degree of wear, represents factors that leads to extension of mounting period of a modular laser system type TOPCON on a leveling machine.

This chapter presents the structure, performance, and the optimal synthesis by numerical simulation of a testing bench designed for TOPCON laser modular systems, which reproduce the operating conditions of the systems set up on the real leveling machines. The device was developed by the aid of a numerical simulation model built in AMESim, is an electro hydraulic servomechanism of position control with feedback by laser. This servomechanism contains two internal control loops for position control: the first loop appears at the level of a servomechanism that simulates the profile of uneven soil, and the

second control loop appears at the level of a tracing servomechanism with laser reference. Dynamic performance obtained by numerical simulation and experimental identification of a TOPCON laser modular system, are in good agreement with those obtained by comparative testing of the same system mounted on a motor grader, in actual operation (Popescu et al., 2008, 2009).
The testing bench developed by INOE 2000-IHP Bucharest, allows the laboratory tuning always needed to be made in order to fit the parameters of laser modular system with the parameters of machine on which are to be mounted. If some malfunction occurs in operation of the machine equipped with a laser modular system, using the testing device one can detect which component of the system no longer provides the functional parameters (laser emitter, laser receiver, electrohydraulic block or electronic block).

2. Laser modular systems made for equipping the ground leveling installations

The leveling technology which uses laser, fig.4, implies a leveling performed by a complex installation, equipped with laser controlled modular system, which is able to perform work from 2 passes, a rough leveling and a fine one at finish, with deviations from the reference plane of max. 2,5 cm on the entire leveled surface with a significant reduction of the tracking, transposition and materialization process during the leveling work.
The modular mechatronic system with laser, electronic and electrohydraulic components, which allows reaching this leveling technology may be mounted on any land leveling equipment whose work bodies, scoops or blades are hydraulically powered. It is conceived as an additional option of the land leveling equipment, which offers to it the possibility of leveling land automatically, without any human error occurrence in what regards precision.

a) b)

Fig. 4. The laser leveling technology: a) mounting the laser modules transmitter and receiver; b) automatic leveling after an optical leveling plan PON performed simultaneously by 6 land leveling equipments UTN

In the classic acception land leveling controlled by laser systems implies a modular system with the following structure.

a. **The laser transmitter** placed in the center of the surface to be leveled above a point with a known quote mark, on a tripod which may be adjusted vertically, emitting a laser beam in its rotation movement. This generates the laser reference plane or the optical refence plane (with programming options for the longitudinal and transversal slope in

the forward direction) which the work body of the equipment will follow during leveling. After setting the slope needed at leveling, the laser transmitter positions itself automatically.

b. **The laser receiver**, whose support is connected to the work body of the land leveling equipment, intercepts the laser fascicle generated by the laser transmitter and sends altimetric information, namely the position of the work body accountable to the laser reference plane, to an electronic control and monitoring module, placed in the cabin of the land leveling equipment.

c. **The electronic monitoring and control module** which conects and amplifies the laser information received, compares it with a prescribed dimension specific for the leveling quote value, finds the error and emits a prompt for cancelling error, towards an electrohydraulic drive system.

d. **The electrohydraulic system** controlled by the electronic module has the role of driving the hydraulic cylinders of the blade for maintaining the work body in the leveling plane set by the leveling project, plane which is parallel with the laser reference plane. In fig.5. are presented two types of land leveling machines equipped with modular systems.

a) b) c)

Fig. 5. Leveller (a) and autograder (b) equipped with laser controlled modular systems (c)

3. Simulating the real operational conditions of the laser module by an original test bench

In fig.6-a is shown the laboratory test bench which simulates the real behavior of the TOPCON laser controlled modular system purposefully created for equipping the automatic land leveling machines in horizontal plane (Popescu et al., 2008). Fig.6-b shows the mode of equalization of the device with the simulation model from AMESim.

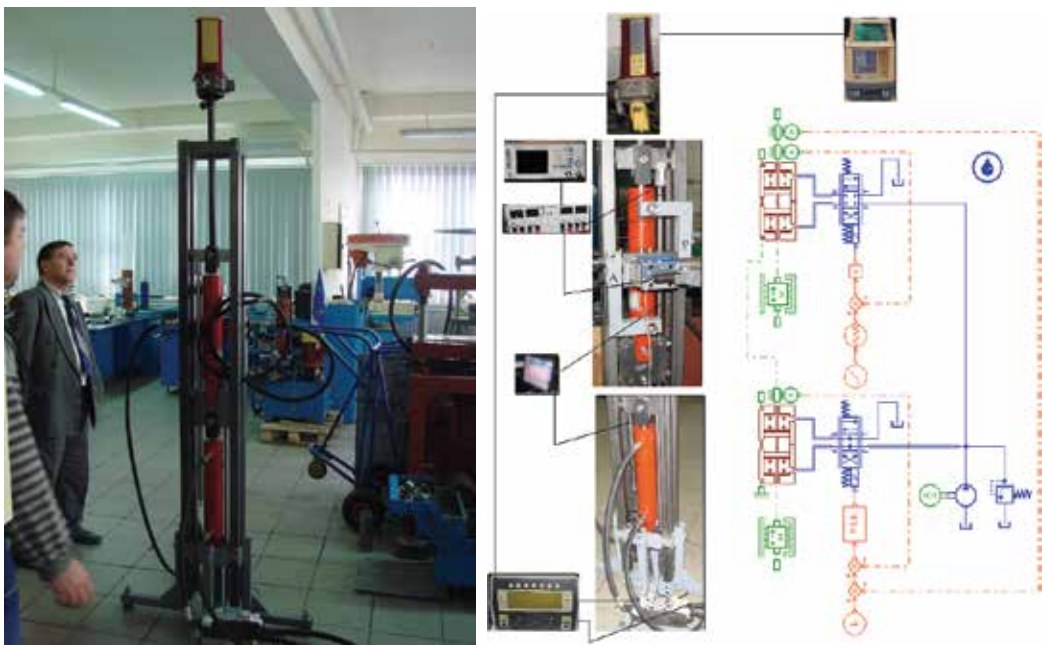
On the rod of the upper hydraulic cylinder is fixed the laser receiver which may move by the action of the upper cylinder or of the bottom cylinder or of both.

The device for testing the laser controlled equipment includes 2 electro hydraulic servomechanisms that simulate the real behavior of the upward downward hydraulic cylinders of the blade of the land leveling machine, and the second - the profile of the land to be levelled.

The first servomechanism contains a hydraulic cylinder similar with that mounted on the machine, supplied from the hydraulic delivery block TOPCON depending on the level of detection of the laser reference plane, generated by a rotary laser transmitter TOPCON.

The second servomechanism consists of a hydraulic servocylinder controlled by a proportional valve with integrated electronics, by means of a data aquisiton board, a PC and the data aquisition software TEST POINT produced by Capital Corporation from USA.

The TOPCON electronic block receives electric signal from the laser receiver, placed on the rod of the upper cylinder of the device. The signal size varies depending on the level of detection of the optical reference plane, generated by the rotary laser transmitter; the input sent to the proportional valve of the TOPCON hydraulic kit is proportional with the detection level. According to this prompt the rod of the bottom cylinder pulls or pushes the body of the upper cylinder in reverse direction to that of deplacement of the cylinder rod.
The upper cylinder is controlled in close loop by means of a servocontroller; a signal generator simulates various profiles for the uneven land. The two inductive transducers of linear deplacement of the cylinders are connected by means of a data acquisition board to a PC using TEST POINT DAS.



a) b)

Fig. 6. Test bench for testing the TOPCON laser controlled modular system: a) general overview; b) the test bench versus AMESIM simulation model

4. Basic mathematical model of the test bench components

A deep understanding of the upper phisical performance needs at least a mathematical modeling and simulation of an electro hydraulic servomechanism.
The simplest nonlinear realistic mathematical model of such a system contains the following equations (Vasiliu &Vasiliu, 2005):

a. The steady-state characteristics of the servovalve main stage (four way, critical centre, spool valve):

$$Q_{SV}(x,p) = c_d A(x)\sqrt{\frac{p_S - P}{\rho}} \tag{1}$$

Here x is the spool displacement from the neutral position; P - pressure difference between the ports of the hydraulic cylinder; A(x) – metering ports surface; c_d – discharge coefficient of the metering ports; p_S - supply pressure (a constant). The above relation can be written in the form

$$Q_{SV}(x,P) = c_d \pi d_s x \sqrt{p_S / \rho} \sqrt{1 - P / p_S} = K_{Qx} x \sqrt{1 - P / p_S} \tag{2}$$

where

$$K_{Qx} = c_d \pi d_s \sqrt{p_S / \rho} \tag{3}$$

is the "flow valve gain".

b. The spool motion equation. The servovalves manufacturers specify for each device the transfer functions adequate to slow, normal and high-speed control process. For slow control process, the servovalve can be regarded as a proportional device, having a single constant - the displacement-current (voltage) gain:

$$K_{xi} = \left. \frac{\partial x}{\partial i} \right|_{x=0} \tag{4}$$

Hence the spool motion follows the input current, i without any lag:

$$x = K_{xi} i \tag{5}$$

For normal control process, a servovalve can be regarded as a first order lag device:

$$\frac{x(s)}{i(s)} = \frac{K_{xi}}{T_{SV} s + 1} \tag{6}$$

The corresponding differential equation is:

$$T_{SV} \frac{dx}{dt} + x = K_{xi} i(t) \tag{7}$$

Here T_{SV} is the servovalve time constant. For high speed control process, we have to consider the servovalve as a second order lag device:

$$\frac{x(s)}{i(s)} = \frac{K_{xi}}{(s / \omega_n)^2 + 2s\zeta / \omega_n + 1} \tag{8}$$

where ω_n is the natural frequency and ζ - damping coefficient.

c. The position transducer equation. The modern inductive position transducers together with their amplifiers behave as first order lag devices; they have a very small time constant, which can be neglected for industrial electro hydraulic control process:

$$U_T = K_T y \tag{9}$$

where K_T is the transducer constant, and y – piston displacement from the null position.

d. The error compensator equation. This stage computes the following error, ε as a difference between the input signal, U_i and the position transducer output, U_T , and applies the PID control algorithm to find the solenoid control voltage, U_c :

$$U_c(s) = \varepsilon(s)K_P[1+1/(sT_i) + sT_d/(\tau s + 1)] \tag{10}$$

e. The servocontroller current generator equation. The current generator of the servo-controller is so fast than it can be regarded as a proportional device:

$$i = K_i U_C \tag{11}$$

where K_i [A/V] is the "medium" conversion factor.

f. The continuity equation. This equation offers the connection between the servovalve flow and the derivative of the pressure drop across the hydraulic cylinder:

$$Q_{SV} = A_p\dot{y} + K_l P + \frac{A_p^2}{R_h}\dot{P} \tag{12}$$

where A_p is the piston area; K_l - leakage coefficient between the motor chambers; R_h - hydraulic stiffness of the motor:

$$R_h = 2\frac{\varepsilon_e}{\mathbf{V}_t}A_p^2 \tag{13}$$

Here ε_e is the equivalent bulk modulus of the oil and $\mathbf{V}_t$ - the total volume of the oil from the hydraulic motor and the connections.

g. The piston motion equation. The pressure force F_p has to cover the load force, usually modelled by a spring force F_e , the inertia of all the moving parts, m_e and the friction force, F_f with different components:

$$m_e\ddot{y} = F_p - F_e - F_f \tag{14}$$

where

$$F_p = A_p P \tag{15}$$

$$F_e = 2\left(K_{e1} + K_{e2}\right)\left(y + y_{0e1}\right) = 2K_e\left(y + y_{0e}\right) \tag{16}$$

The friction force has mainly a static component, F_{fs} and a viscous one, F_{fv}:

$$F_{fs} = F_{fs0}sign\,\dot{y} \tag{17}$$

$$F_{fv} = K_{fv}\dot{y} \tag{18}$$

The main non-linearity in the above mathematical modeling is included in the servovalve main stage. A linear solution can be obtained using a linear form of the steady-state characteristics of the servovalve main stage,

$$Q_{SV} = K_{Qx}x - K_{QP}P \tag{19}$$

The results supplied by the linear model are useful for estimating the stability only. For high amplitude input signals the designer has to use the numerical simulation. Some simulation languages are widely used for practical purposes. Two of them are available for any engineering activity: SIMULINK (The Math Works Inc., 2007) and AMESIM (LMS Imagine, 2009). The „building" of a simulation network in SIMULINK needs a lot of work for using general purpose „icons", but the toolboxes devoted to sistems synthesis are very effective. The fig. 7 contains the simulation network of the above electro hydraulic servomechanism.

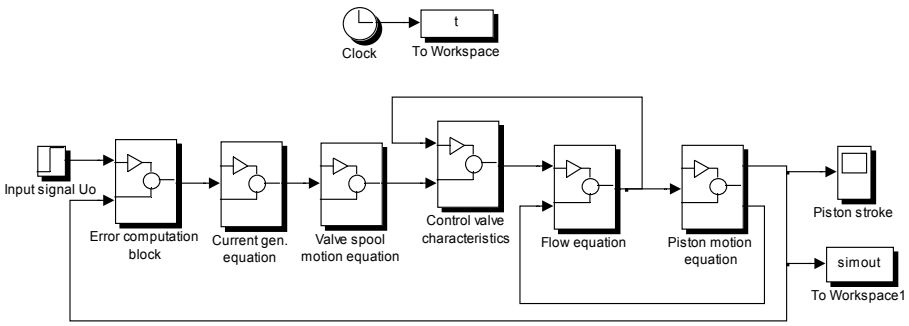


Fig. 7. Simulation network of an electro hydraulic servomechanism in SIMULINK

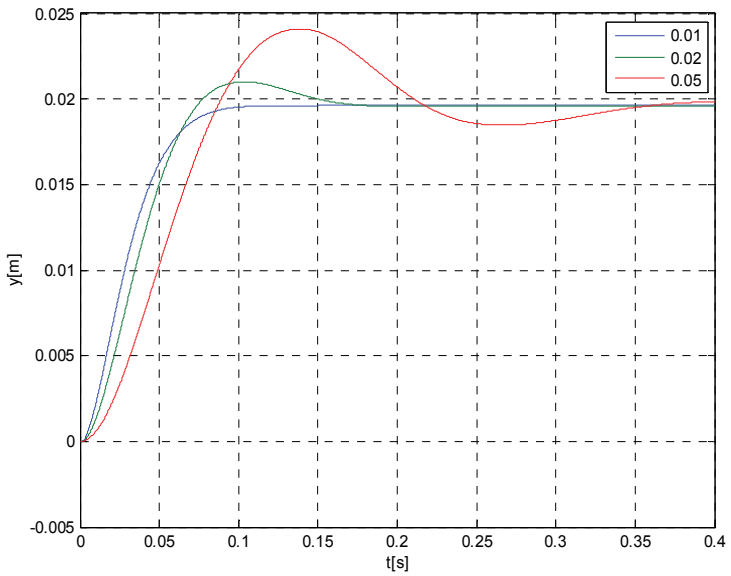


Fig. 8. Small input step response of an electro hydraulic servomechanism simulated by SIMULINK for three values of T_{SV}: 0.01 s, 0.02 s, and 0.05 s

Figure 8 presents the response of the servomechanism for small step inputs, and three values of the servovalve time constant, T_{SV}. Using a high speeed servovalve one can obtain an overall small time constant of about 0.045 s. The increase of the servovalve time constant spoils the system dynamic performance and can generate steady state oscillations. A long series of experiments were performed by (Calinoiu, Vasiliu & Vasiliu, 1998) in order to find

the difference between the theoretical dynamic behaviour and the real one for a servomechanism using a Bosch NG10 direct drive servovalve (DDV). There is a good agreement between the simulated and measured results, the time constant having practically the same value for both cases. The Bode diagram (fig. 9) shows a good dynamics even for a high spring load. On the same diagram the transfer function identified by IDENTIFICATION TOOLBOX from MATLAB is specified. The computed transfer function and the measured one are nearly the same.

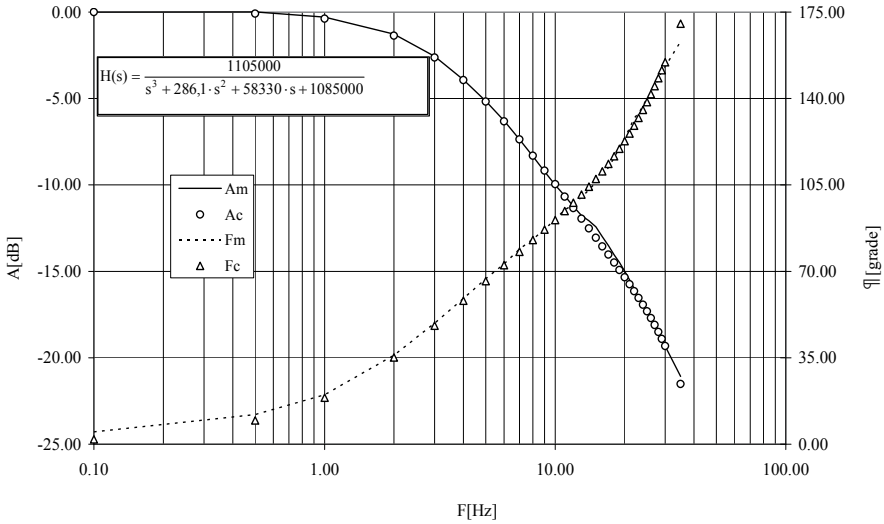


Fig. 9. Bode diagram of an electro hydraulic servomechanism (identification by MATLAB)

The simulation model of the test bench can be assemblied by two SIMULINK models as in fig. 7, but the capabilities of AMESIM are very useful for the quick design and optimization.

5. AMESIM design facilities

5.1 Overview

Many different modeling and simulation software packages were created to perform studies in the fields of automobile, aerospace, robotics, offshore and general hydraulics engineering but none offered the full range of capabilities needed. There were deficiencies in the numerical capabilities, in the graphical interface and in the general modeling concept. The AMESim package was developed to overcome these limitations by Michel Lebrun and Claude Richards from Societe Imagine (FRANCE), starting from 1988. This section gives a description of the technical features, which were central objectives in the design of the software, and some examples of typical applications.

The main aim of the AMESim is "To create Good Models without Writing a Single Line of Code" (Lebrun & Richards, 1997). An important prerequisite of the basic element library is the creation of extremely well tested, reliable and reusable submodels that a user can employ with complete confidence (LMS IMAGINE SA, 2009). The writer of the basic element library must be competent in all the modeling skills. However, the user of the basic element library is relieved of the need to write code and formulate the mathematics. Understanding of the details of the physics is not needed but decision on assumption is

necessary which imply some knowledge of physics. Understanding of the engineering system and an ability to interpret results is still important. Experience in training design office staff to use of the basic element library suggests that it is learnt very rapidly.
AMESim is using the multiport approach. In the signal port approach of a numerical simulation environment, a single value or an array of values are transferred from one component block to another in a single direction. This is fine when the physical engineering system behaves in the same way such as with a control system. However, problems arise when power is transmitted. This is because modeling of components that transmit power leads to a requirement to exchange information between components in both directions. In order to use a signal port approach in this situation, two connections must be made between the components where physically there is only one. This leads to a great complexity of connections and means that even very simple models involving power transmission appear complex and unnatural. In contrast to the signal port approach, with the multiport approach, a connection between two components allows information to flow in both directions. This makes the system diagram much closer to the physical system.

5.2 Numerical performance

The analysis of the steady state and dynamic behavior of an engineering system leads to a mathematical model of the system. This is in the form of algebraic, ordinary differential and partial differential equations. More recently, differentialalgebraic equations are also used to model the system. The role of simulation software is to provide an environment in which this model can be solved efficiently. For models with large numbers of partial differential equations, there are specialist packages such as those for computation fluid dynamics. Such software is used for detailed analysis of individual components of a system. However, it is often necessary to simulate a completely engineering system or a subsystem of it. The concept of the virtual prototype, in which physical prototypes are replaced by mathematical computer models, makes simulation of this type vital. In this case, it is normal to reduce any partial differential equations to ordinary differential equations. This leads to models with either ordinary differential equations (odes) or differential algebraic equations (daes). Many general and specialized simulation software packages are available for solving such systems of equations. Models arising from engineering systems vary greatly in their character. Thus the equations of the model can be: linear, non-linear, numerically stiff i.e. with very small time constants compared with the overall simulation period, oscillatory, continuous, discontinuous. A large variety of numerical integration methods can be employed to solve such problems. Traditionally the user of simulation software is presented with a menu of typically seven methods from which a choice must be made.

5.3 Direct access graphical user interface

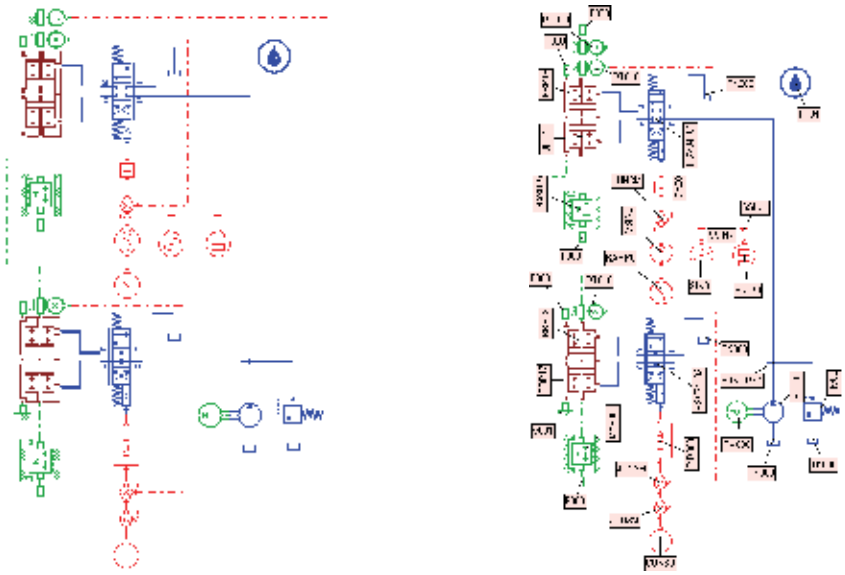
Many older simulation packages were developed before modern graphical user interfaces were available. The only graphical facilities provided were for producing simple plots of results. The suppliers of these packages have had to introduce new graphical preprocessing facilities to build the system. More modern software has been designed from the start with a full graphical user interface. Whenever possible, icons for components were based on internationally recognized standard symbols. Thus for hydraulic systems icons are based on CETOPS symbols. Where there are no such standardized symbols, icons are constructed which can be instantly recognized by engineers working in the field.

Throughout the simulation, process the system diagram is displayed. Thus for example when parameters are changed for a particular component, the user points at the icon in question and clicks the mouse button. This produces a menu of items that may be changed. Similarly to plot graphs of results, the user points at the component and clicks the mouse button to produce a menu of items associated with the component that may be plotted.
The possibility of quick high level technical developments as ABS, EBS, common rail multipoint injection systems, electro hydraulic automatic transmissions, self tuning hydraulic and pneumatic suspensions, hydraulic power steering, fly-by-wire systems and many others (Mare & Cregut, 2001). Companies like AEROSPATIALE, MATRA, BOSCH, FERRARI, DAIMLER-CRIYSLER, GENERAL MOTORS, etc. are currently using this modeling and simulation software for future developments. Academic training programs are now developed in different countries, for teaching the software in the terminal years (Vasiliu & Vasiliu, 2005), and for applied researches (Vasiliu, et al., 2003).

6. Numerical simulation and experimental identification of the laser controlled modular system by AMESim

6.1 Modelling the test bench

For the numerical simulation of the laser controlled modular system it was used the simulation in AMESim, namely the model shown in fig.10. All the components of the simulation model are based on mathematical models of differential equations, validated by practice and the method of numeric integration of the differential equations is chosen automatically. If the model is not correct or the inner and outer parameters are not properly determined, the program does not work, cause the system of differential equations is incompatible or undetermined.



a) b)

Fig. 10. Model of simulation in AMESim for a TOPCON laser controlled modular system mounted on a testing device: a) simulation model; b) model components

The simulation model represents an electrohydraulic servomechanism for adjusting the position with laser reaction. It includes 2 inner adjustment loops and an outer loop. The first inner loop is set at the level of the hydraulic servomechanism of simulation for uneven land which is excited at entry with rectangular, sinusoidal signals, constant and variable. The second inner loop is set at the level of the servomechanism of monitoring with laser control which is similar to the TOPCON laser controlled modular system. The outer loop of regulation is done between the exit of the first servomechanism and the entry of the second.

6.2 Numerical simulation experiments

In fig.11…17 are shown some of the significant numeric simulations. In fig.11 the servomechanism generating profiles of the uneven land receives a rectangular input with an amplitude of 0,14 m and a frequency of 0,05 Hz in a range of 50 s. The red curve 1 represents the desplacement of the rod of the generator servocylinder [m], and the green curve 2 represents the rod deplacement of the monitoring servocylinder rod and the body of the generator servocylinder [m].

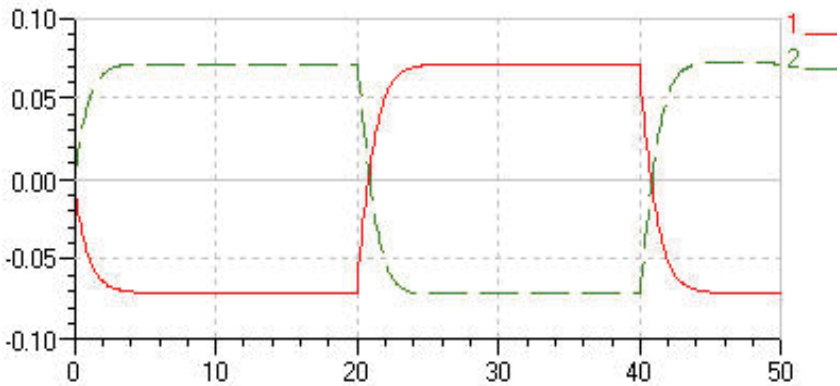


Fig. 11. The answer of the laser monitoring servomechanism at exciting the servomechanism which generates profile with rectangular signal

By the algebraic sum of the graphics from fig.11 results the curve 3 from fig.12. In the terminology related to the operation of automatic land leveling after an horizontal plane curve 3 represents the deviations of the profile of the levelled land from the optical horizontal reference plane. These are present only in the zone of stage jumping last 2 s and have a max.value of 0,01 m.

In fig.13 the servomechanism generating the profile of uneven land is excited with a constant sinusoidal signal with an amplitude of 0,14 m and a frequency of 0,05 Hz lasting 50 s. The meaning of the curves 1 and 2 is the same with that from fig.11.

By the algebraic sum of the graphics from fig.13 it results the curve 3 from fig.14 with the same meaning as that from fig.12. The errors are negligible with max.values below 0,002 m.

In fig.15. is shown a method for emitting in AMESim a sinusoidal signal with variable amplitude and frequency: over the sinusoidal signal with variable frequency and constant amplitude is superposed a ramp signal after this the 2 signals being composed. For the component signals there is a model in AMESim but for the composed signal not.

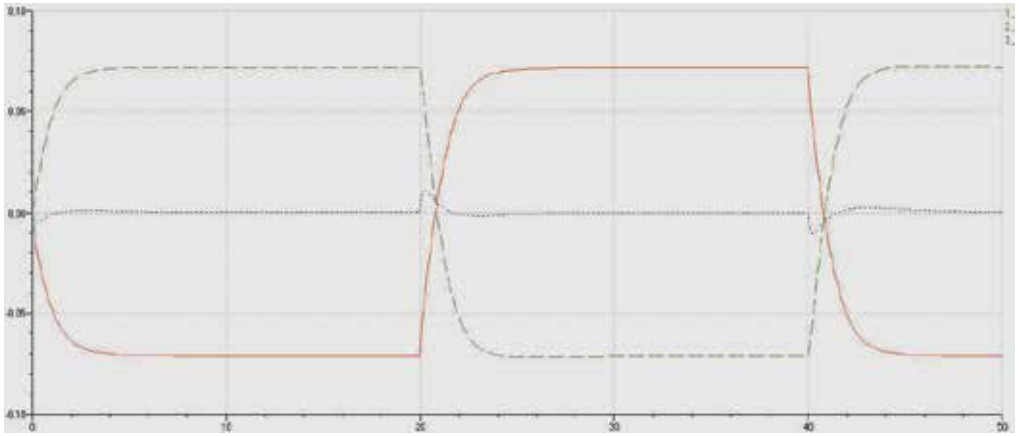


Fig. 12. Deviation profile of the leveled land from the optical reference plane

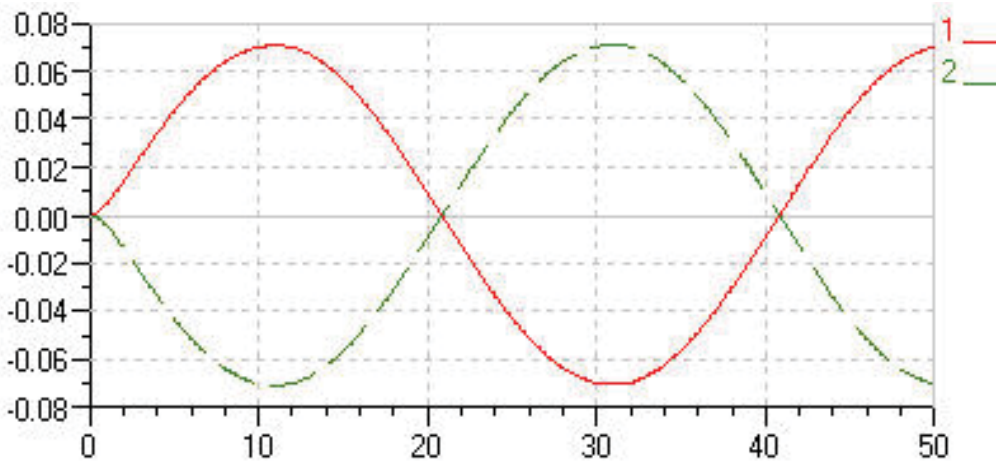


Fig. 13. The answer of the laser monitoring mechanism for a constant sinusoidal input

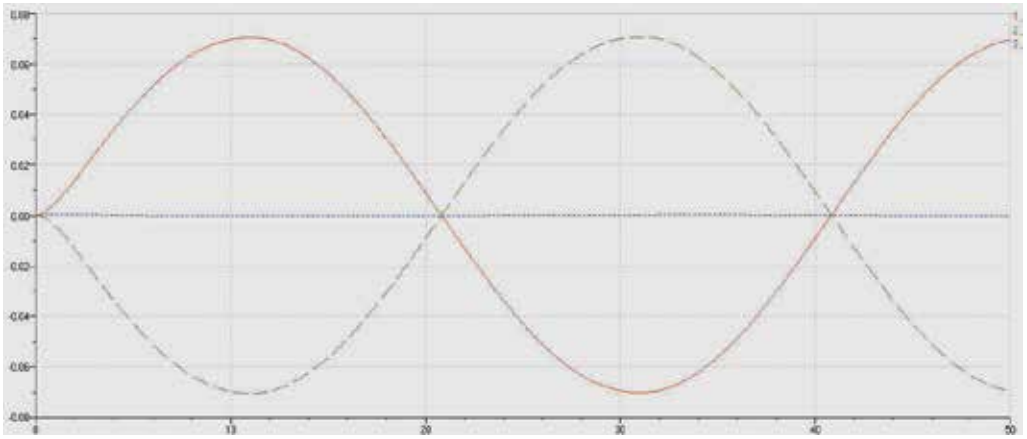


Fig. 14. Deviation profile leveled land from the optical reference plane

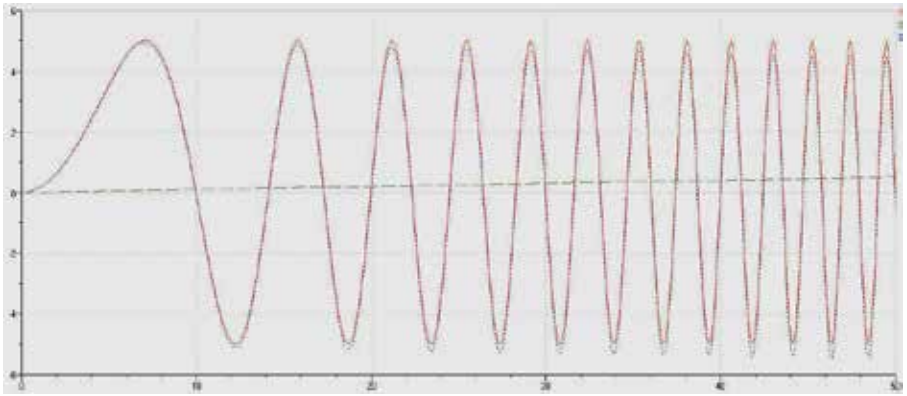


Fig. 15. The formation of a sinusoidal imput signal with variable frequency and amplitude

The meaning of the curves from fig.15. is the following: 1- sinusoidal signal with variable frequency, max. frequency 0,5 Hz and amplitude 0,1 m; 2 – ramp signal; 3- sinusoidal signal with variable frequency and amplitude.

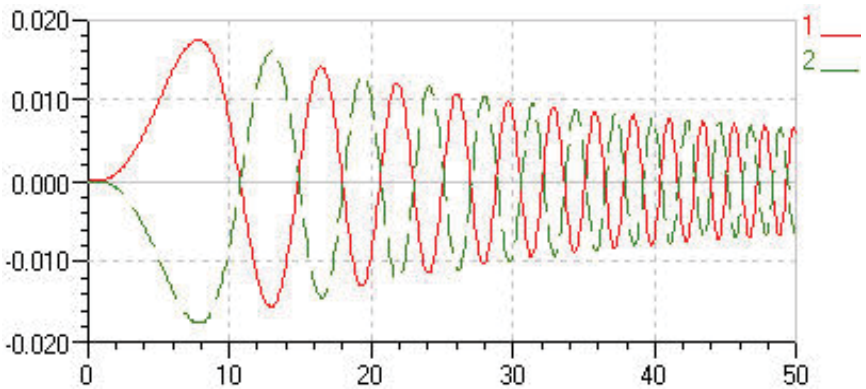


Fig. 16. The answer of the laser monitoring servomechanism at exciting the servomechanism generator of profile with variable sinusoidal signal

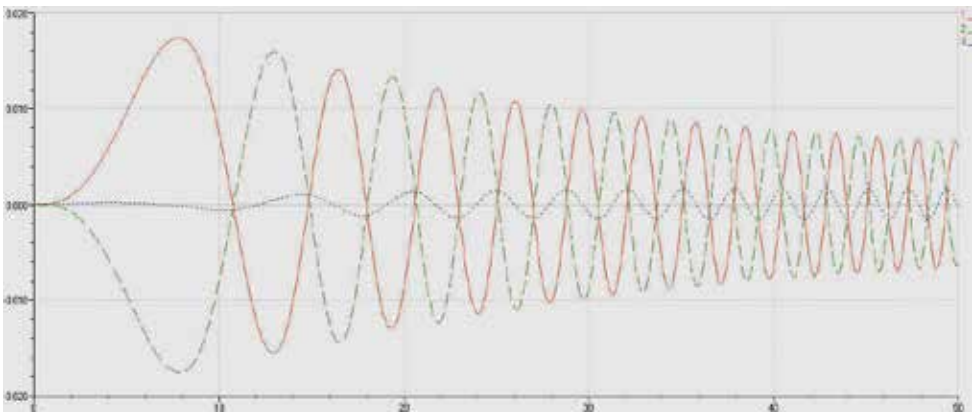


Fig. 17. Deviation leveled land profile from the optical reference plane

In fig.16 the servomechanism generator of uneven land profile is excited with a variable sinusoidal signal of the shape shown in fig.15 with an amplitude of 0,14 m and a frequency of 0,05 Hz lasting 50 s. The meaning of the curves 1 and 2 is the same like that shown in fig.11. By the algebraic sum of the graphics from fig.16 it results the curve 3 from fig.17 with the same meaning like that presented in fig.12. The errors are negligeable with frequencies below 0,8 Hz and a max.value of the deviation of 0,004 m.

6.3 Fine tuning the parameters of PID controller

The modern fluid control systems are using hybrid tuning alghoritms as Fuzzy - PID error compensators (Popescu et al., 2009). The high degree of nonlinearity of these systems leads to the wide use of modeling and simulation techniques for obtaining the tuninig parameters by a virtual testing system. This testing manner offers a strong costs cut, and a useful reduction of the real experimental test. After 20 years of intensive development of the symbol libraries in different engineering fields, AMEsim became an efficient tool for solving different applications of the fluid control systems. The case presented in this paper intends to offer a model of developping new applications of the electro hydraulic systems by this tool. The authors created both the laboratory model of the electro hydraulic control system, and the real system set up on a modern ground leveling machine. The comparison between the static and dynamic performances of the real system is found in good agreement.

To tune a controller means to find the parameters of an given structure, of a settled degree, so that to achieve from the resulted system a behavior as close as possible to the desired one. In practice the most frequently used regulators are of type P, PI, PD and PID which calculate the $u(t)$ command according to the following relations: (1), for a **P**: regulator: proportional; (2), for a **PI**: compensator proportional, integral; (3), for a **PD**: regulator proportional, derivative; (4), for a **PID** regulator proportional, integral, derivative, where: K_P – constant of the proportional part (gain), K_I – constant of the integral part, K_D – constant of the derivative part.

$$. u(t) = K_P \cdot \varepsilon(t) . \tag{20}$$

$$u(t) = K_P \cdot \varepsilon(t) + K_I \cdot \int \varepsilon(t) dt \tag{21}$$

$$u(t) = K_P \cdot \varepsilon(t) + K_D \frac{d\varepsilon(t)}{dt} \tag{22}$$

$$. u(t) = K_P \cdot \varepsilon(t) + K_I \cdot \int \varepsilon(t) dt + K_D \frac{d\varepsilon(t)}{dt} . \tag{23}$$

PID type controllers are used for the error signal in hydraulic rapid servomechanisms. Component **P** amplifies the error, develops a higher-speed system, but it can't cancel the stationary error; component **I** removes the stationary error, but it destabilizes the system, while component **D** stabilizes the system. The last generation of control algorithms are based on the real time simulation of the systems.

The simulation model in AMESim (fig.10a) represents a hydraulic servomechanism for position control with one external feedback by laser and two internal feedbacks, arising at the level of the two included servomechanisms, as follows. The upper servomechanism, that simulates the profile of the uneven land, and the lower servomechanism, a tracing one, that actuates the blade of the levelling machine in a vertical plane.

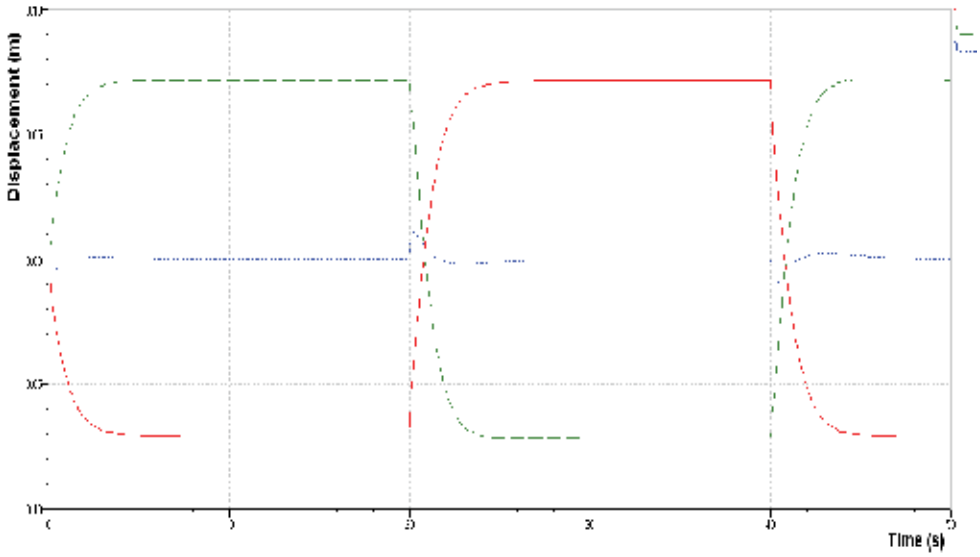


Fig. 18. The response of the tracing servocylinder when the servocylinder that simulates the land profile is excited by a rectangular signal

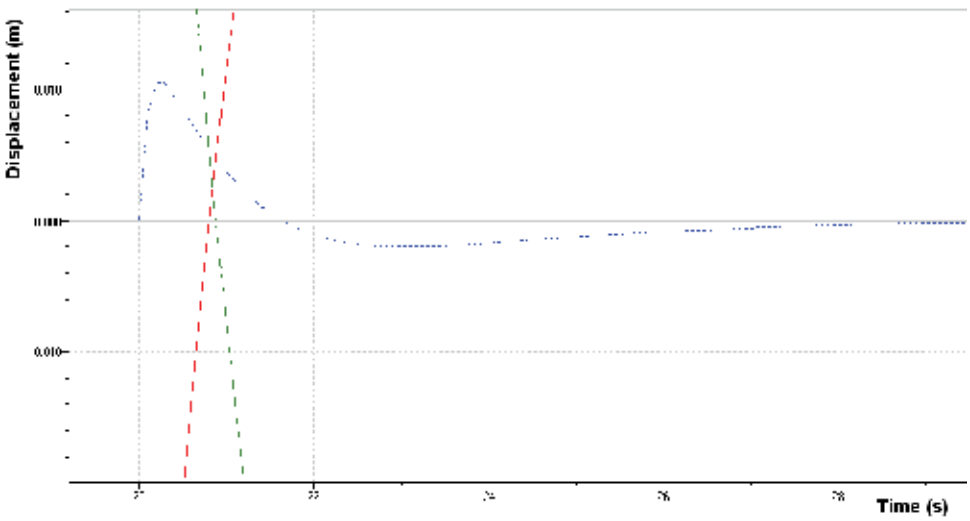


Fig. 19. The maximum value of the deviation of the leveled land from the optical reference plane

The hydraulic servo cylinder of the upper servomechanism has a mobile body, while the one of the lower servomechanism has a fix body. The first internal feedback loop arises between the displacement transducer of the cylinder with mobile body and the upper comparator of the simulation model. The second internal feedback loop arises between the displacement transducer of the cylinder with fix body and the internal comparator of the simulation model. The external feedback loop arises between the displacement transducer placed in the upper side of the model and the comparator placed in its lower side. The

above configuration can be a fair representation of the true system, included in the frame of the levelling machine. The servomechanism that simulates the profile of the uneven land is excited by a rectangular signal with amplitude of 0.140 m and frequency of 0.025 Hz. In fig.18 three curves are set: curve1 – variation of displacement over time of the servocylinder that simulates the profile of the uneven land; curve2 - variation of displacement over time of the tracing servocylinder, that actuates the blade of the navvy machine in vertical plane; curve3 – the amount of the two displacement values, which is the variation over time of the deviation of the uneven land from the optical reference plane.

In these conditions the maximum value of the deviation of the leveled land from the optical reference plane is 0.01m, fig. 19.

6.3.1 Optimizing parameter KP

Running the application in AMESim is repeated, this time canceling parameters K_I and K_D and selecting five values for parameter K_P, according to the settings in "Batch Control Parameter Setup" box, fig. 20.

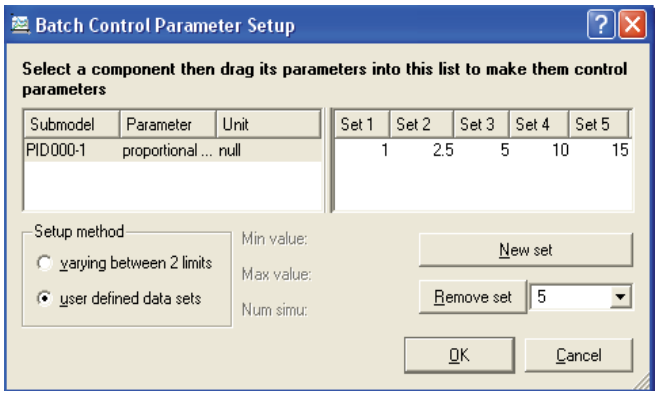


Fig. 20. Setting values for parameter K_P

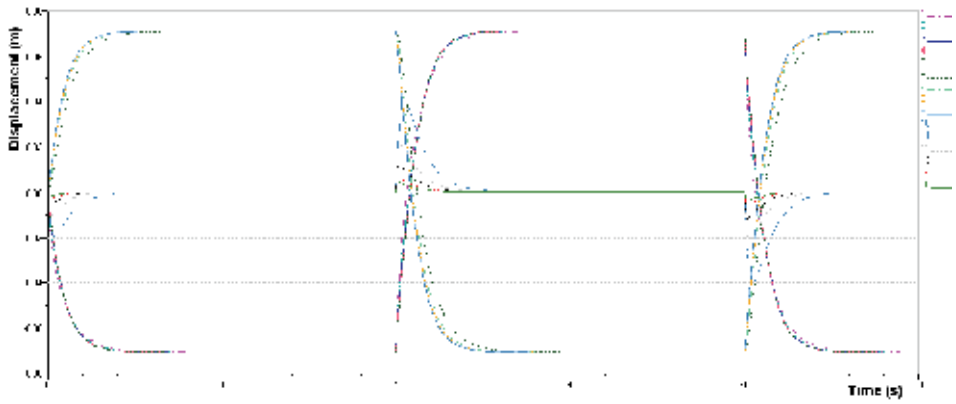


Fig. 21. Influence of the variation of parameter K_P upon the dynamics of the tracing servomechanism

In "Plot manager" box, there are shown the curves resulted when running the application in *Batch* mode, corresponding to five different values of parameter K_P. These curves represent: curve1...curve5 – variation over time of the displacement of the servocylinder that simulates the profile of the uneven land; curve 6...curve 10 - variation over time of the displacement of the servocylinder that actuates the blade of the navvy machine; curve11...curve15 - variation over time of the deviation of the leveled land from the optical reference plane. In fig. 21 is shown the influence that the variation of the parameter K_P has upon the dynamics of the tracing servomechanism when exciting the servomechanism that simulates the profile of the uneven land by a rectangular signal with amplitude of 0.140 m and frequency of 0.025 Hz.
In fig. 22 is shown one detail of the variation over time of the amount of the displacement values of the two servocylinders, when applying the settings in fig. 20. One can notice an increasing dynamics of the tracing servocylinder, in accordance with the increase of the value of parameter K_P.

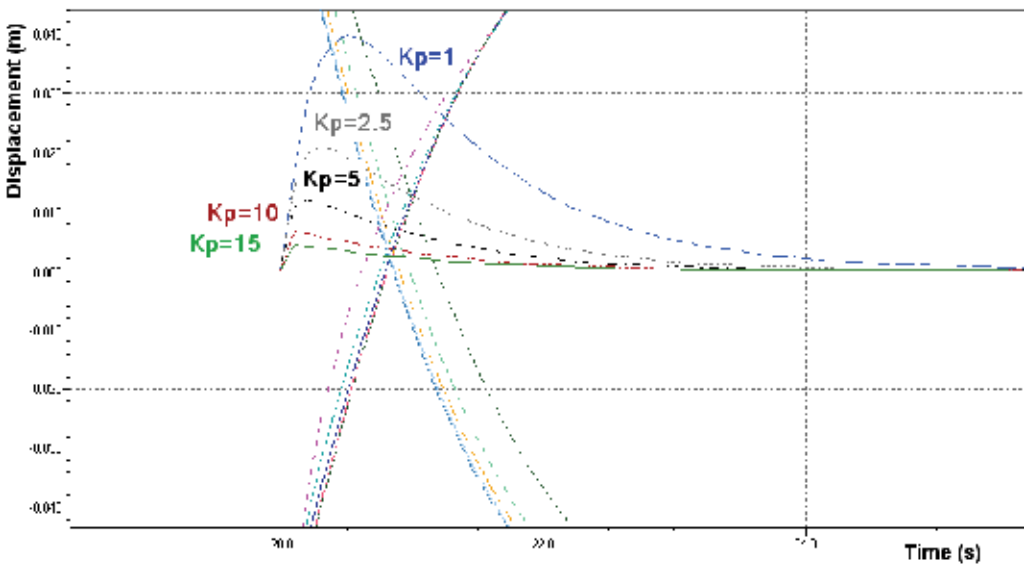


Fig. 22. Variation in the deviation of the profile of leveled land from the reference plane, depending on variation of parameter K_P

6.3.2 Optimizing parameter KI

Running the application in AMESim is repeated, this time canceling parameters K_P and K_D and selecting five values for parameter K_I, according to the settings in "Batch Control Parameter Setup" box. In fig. 23 is shown the influence that the variation of parameter K_I has upon the dynamics of the tracing servomechanism when exciting the servomechanism that simulates the profile of the uneven land by a rectangular signal with amplitude of 0.140 m and frequency of 0.025 Hz. In fig. 24 is shown one detail of the variation over time of the amount of the displacement values of the two servocylinders, when applying the settings K_I=0.5; K_I=1; K_I=2; K_I=4; K_I=8. One can notice that the stationary error in the tracing servomechanism is removed faster at a higher value of parameter K_I.

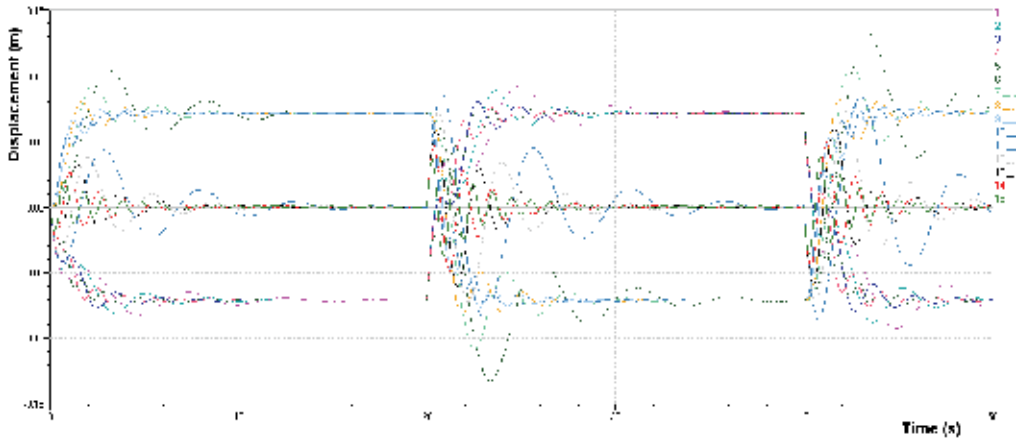


Fig. 23. Influence of the variation of parameter K_I upon the dynamics of the tracing servomechanism

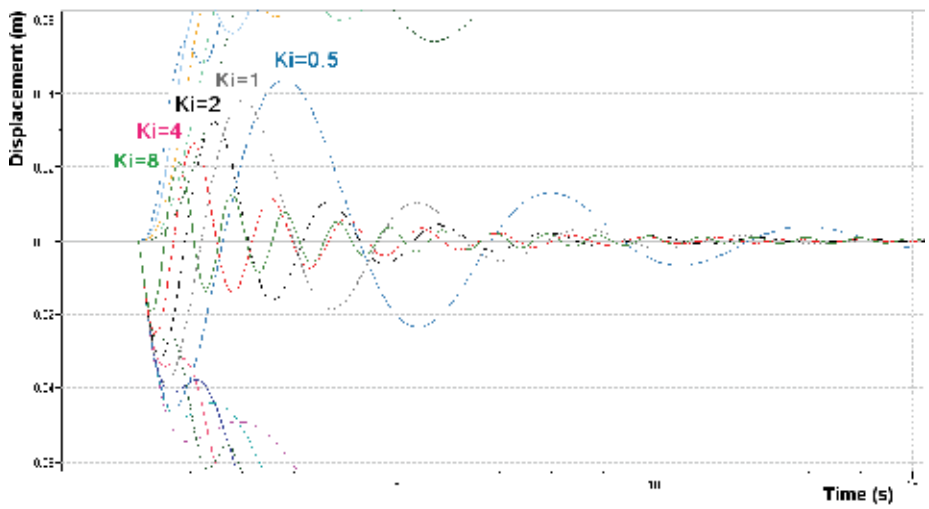


Fig. 24. Variation in the deviation of the profile of leveled land from the reference plane, depending on variation of parameter K_I.

6.3.3 Optimizing parameter KD

Running the application in AMESim is repeated, this time setting parameters K_P=1 ; K_I =0.5 and selecting five values for parameter K_D, according to the settings in "Batch Control Parameter Setup" box. In fig. 25 is shown the influence that the variation of parameter K_D has upon the dynamics of the tracing servomechanism when exciting the servomechanism that simulates the profile of the uneven land by a rectangular signal with amplitude of 0.140 m and frequency of 0.025 Hz. In fig. 26 is shown one detail of the variation over time of the amount of the displacement values of the two servocylinders, when applying the settings K_D=0.1; K_D=0.2; K_D=0.4; K_D=0.8; K_D=1.2. One can notice that the stabilization in the tracing servomechanism is attained faster at a lower value of parameter K_D.

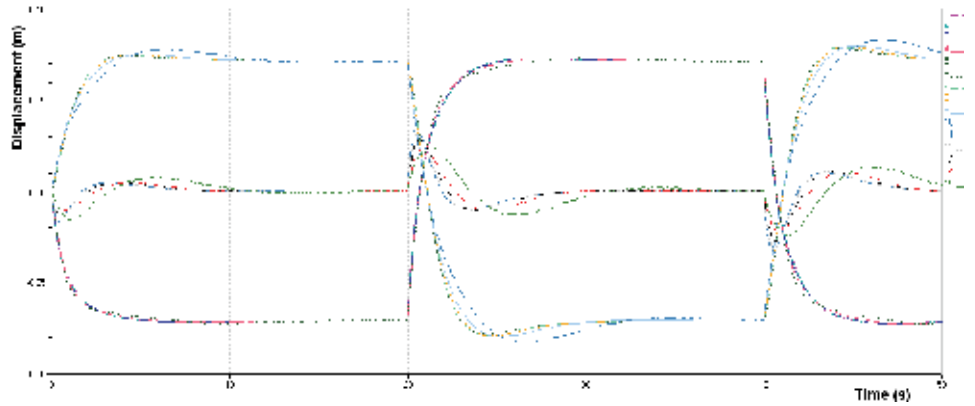


Fig. 25. Influence of the variation of parameter K_D upon the dynamics of the tracing servomechanism

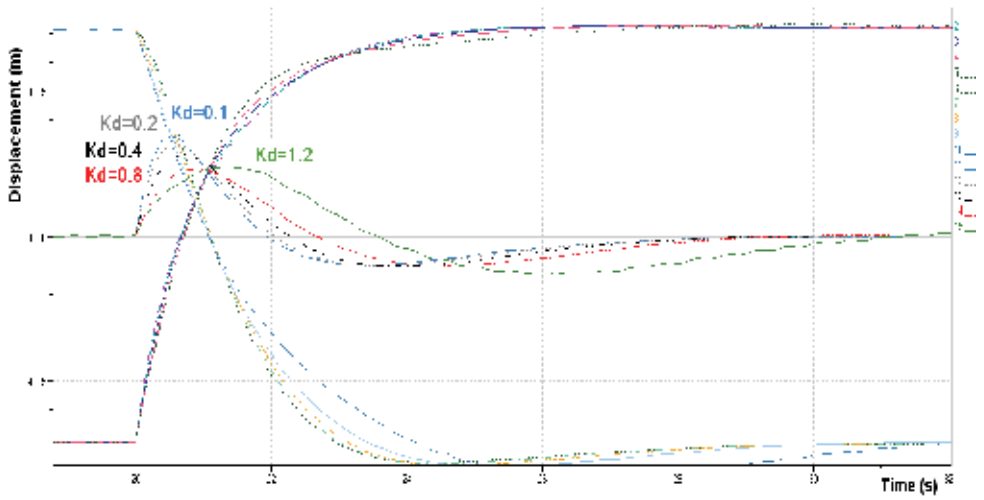


Fig. 26. Variation in the deviation of the profile of leveled land from the reference plane, depending on variation of parameter K_D

6.3.4 Optimizing global parameter K(KP , KI , KD)

Running the application in AMESim is repeated, this time selecting five set of values for parameters K_P , K_I and K_D , according to the settings in "Batch Control Parameter Setup" box. In fig. 27 is shown the influence that the variation of parameter $K(K_P , K_I , K_D)$ has upon the dynamics of the tracing servomechanism when exciting the servomechanism that simulates the profile of the uneven land by a rectangular signal with amplitude of 0.140 m and frequency of 0.025 Hz.

In fig. 28 is shown one detail of the variation over time of the amount of the displacement values of the two servocylinders, when applying the settings: K1(15, 8, 0.1); K2(10, 4, 0.2); K3(5, 2, 0.4); K4(2.5, 1, 0.8); K5(1, 0.5, 1.2). One can notice that the optimal dynamics and stability of the tracing servomechanism is obtained when PID controller has the global parameter K1(15, 8 ,0.1), where: K_P =15 , K_I =8 and K_D =0.1.

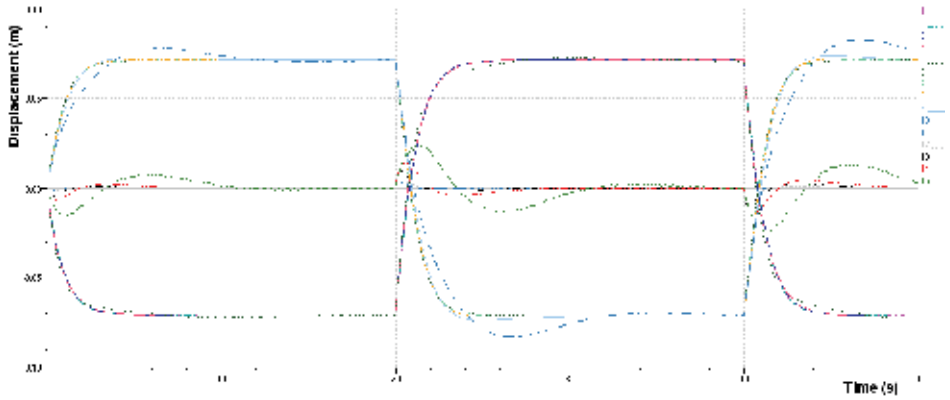


Fig. 27. Influence of the variation of parameter $K(K_P , K_I , K_D)$ upon the dynamics of the tracing servomechanism

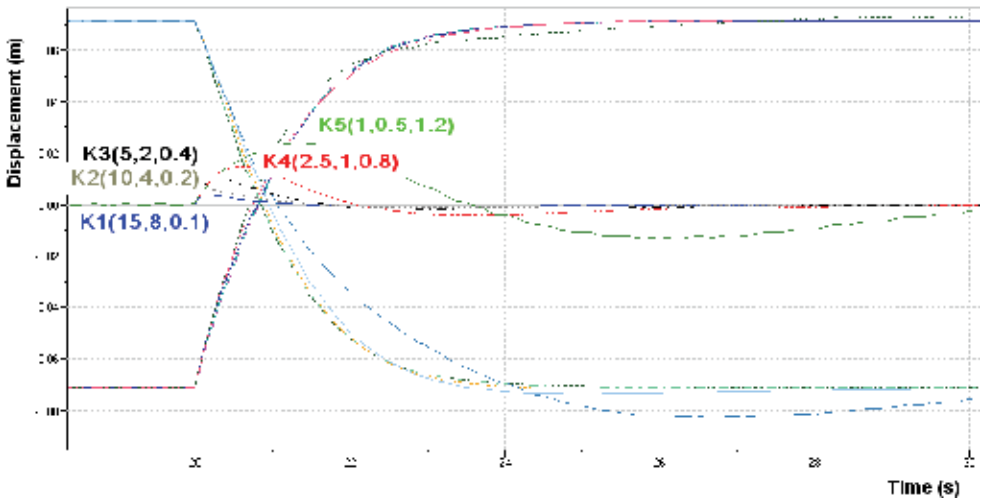


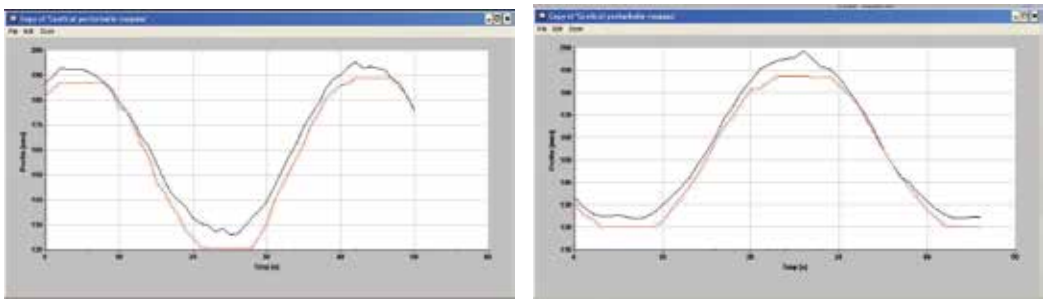
Fig. 28. Variation in the deviation of the profile of leveled land from the reference plane, depending on variation of parameter $K(K_P , K_I , K_D)$

6.4. Experimental identification

The results of the experimental identification of the TOPCON laser controlled modular system mounted on test devices are shown in fig. 29…32.

In fig. 29-a is shown the dynamics of the laser control hydraulic monitoring system when at the input of the hydraulic mechanism generator of uneven land profiles is applied a constant sinusoidal signal with a frequency of 0,025 Hz and an amplitude of 0,072 m. The test duration was 50 s and it proved a proper dynamic of deplacement of the monitoring servosystem (in red) towards the generator of uneven land profile (in black)

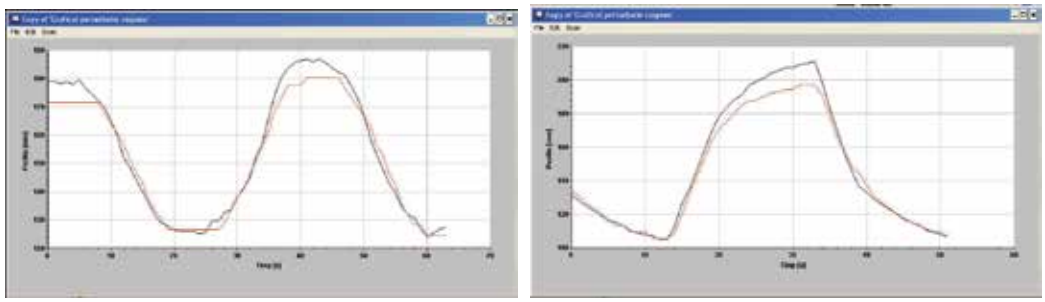
The graphics from fig. 29-b was obtained by repeating the test with the same frequency of the sinusoidal signal of excitation 0,025 Hz but with a higher amplitude 0,080 m. The test took 46 s and the results show a proper behavior of the monitoring servomechanism with laser control.



a) b)

Fig. 29. The answer of the laser monitoring mechanism at the excitation of the servomechanism generator by a constant sine input

In fig.30-a is shown the dynamic of the monitoring hydraulic servosystem with laser control, when at the entry of the hydraulic servosystem generating uneven land profiles it is applied a constant triangle signal with a frequency of 0,025 Hz and an amplitude of 0,060 m which takes 63 s. The test proves the proper work of the laser controlled servomechanism.
In fig. 30-b is shown the dynamic of the hydraulic servosystem with laser control when at the entry of the hydraulic servomechanism generator of uneven land profiles is applied a constant rectangular signal with a frequency of 0,025 Hz and an amplitude of 0,105 m. The test took 51 s.



a) b)

Fig. 30. The answer of the laser control monitoring mechanism at the excitation of the servomechanism generator of profile with: a) triangle input; b) rectangular input

At all tests presented above in fig. 29, and fig. 30 the inductive transducers of lineary displacement of the hydraulic cylinders were set in such a way that the 2 graphics are superposed for noticing easily the dynamic behavior of the hydraulic servomechanism with laser control.
For the test from fig. 31 which uses as excitation signal a constant sine one the inductive transducers of linear displacement of the hydraulic cylinders was set so that they can offer information regarding the real direction of displacement of the cylinders.
In fig. 31 is shown the dynamic of the hydraulic system with laser control when at the entry of the hydraulic mechanism generator of uneven land profiles is applied a constant sinusoidal signal with a frequency of 0,020 Hz and an amplitude of 0,120 m. The test took 115 min.

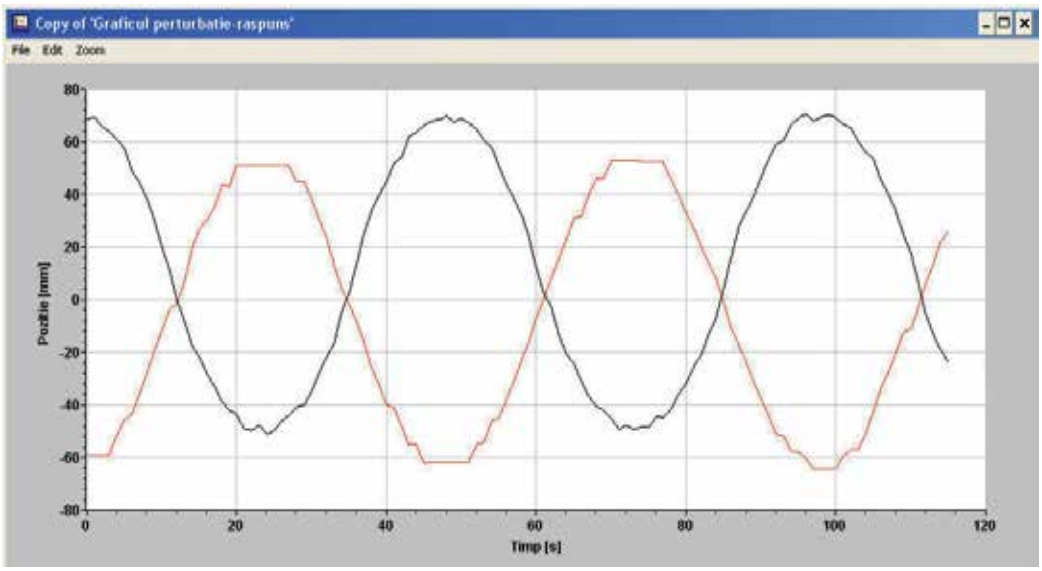


Fig. 31. The answer of the laser monitoring servomechanism at the excitation of the servomechanism generator of profiles with constant sinusoidal signal

In fig. 32 is shown the dynamic of the hydraulic monitoring system with laser control at the excitation of the mechanism generator of variable sine signal with a frequency of 0,010…0,100 Hz and an amplitude of 0,115…0,034 m. The test took 694 s.

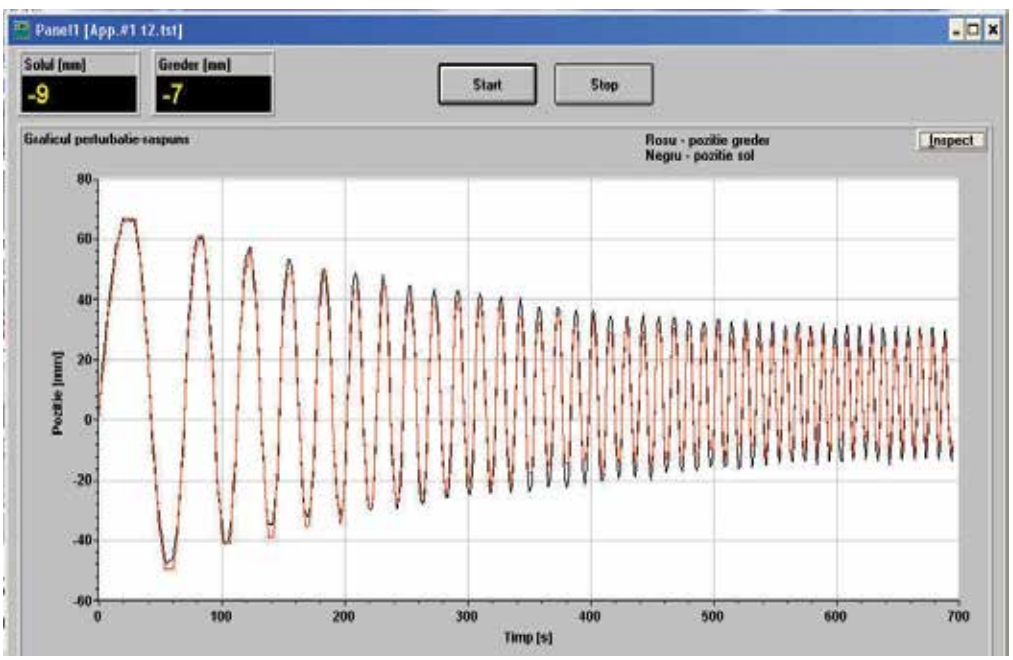


Fig. 32. The answer of the laser monitoring servomechanism at the excitation of the servomechanism generator of profiles with variable sinusoidal signal

Sistematic simulations gave the optimal parameters of the PID controller: K_P =15, K_I =8 s, and K_D =0.1 s (fig.33). The minimum value of the deviation of the leveled land from the optical reference plane is less than 0.004 m (fig. 34). This value is 2.5 times lower than the one resulted from the first running of the simulation model (fig. 17).

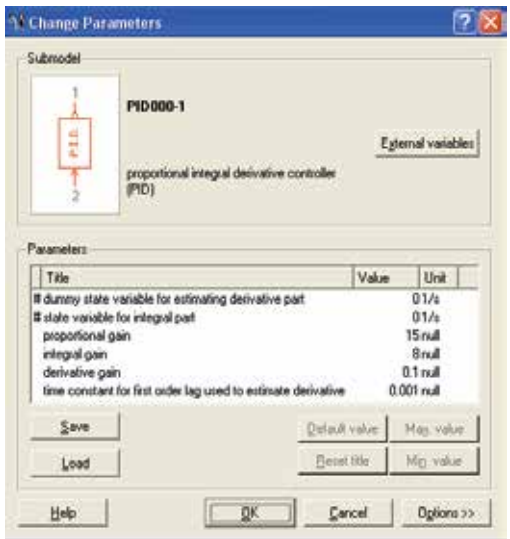


Fig. 33. Setting optimal parameters for a PID controller

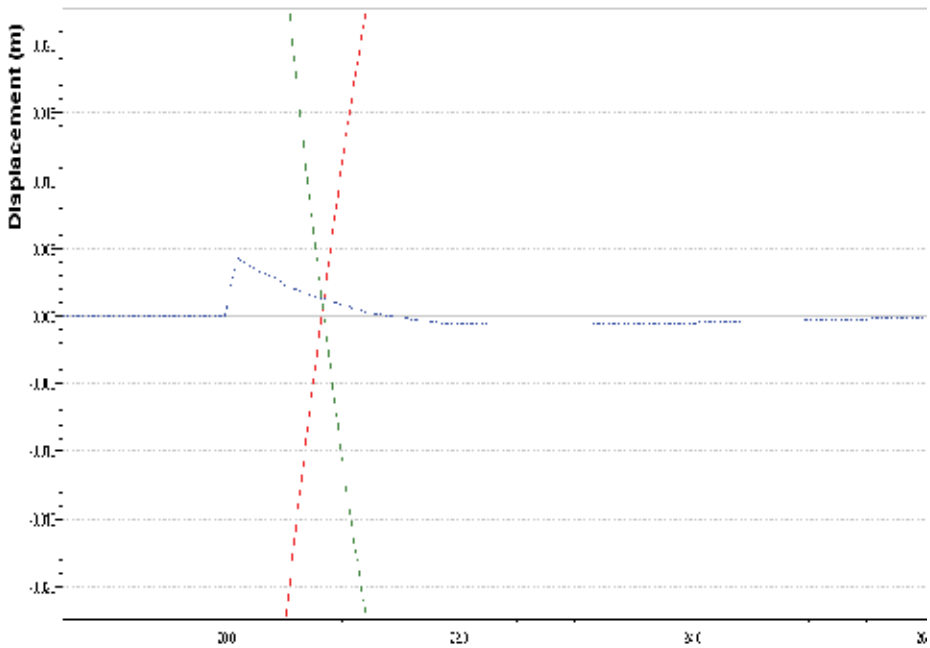


Fig. 34. Maximum optimized value of the deviation of the leveled land from the optical reference plane

7. Conclusions

The laser leveling of the land layers laid down when making a dam or a land dikes from the hydropower stations represents a safe and efficient solution for providing optimum breadth with maximum errors of about 2,5 cm on the entire surface of the laid layer. This kind of leveling performed before compaction of each land layer provide a proper and homogenous density of the dam and represents the optimum solution for reducing infiltrations and avoiding the falling of the crowning which may lead to water flood like is shown in fig.2.
The laser controlled modular systems like TOPCON or similar ones are not standard facilities in civil engineering companies not even for the most modern land leveling machines, but they can be mounted on any kind of hydraulic powdered land leveling machine, no matter of the degree of wear or origin.
The set up of these kind of equipments with laser control systems like TOPCON which appeared in the last decade in Romania is performed by specific trained personnel, and not by the manufacturers of the leveling machines.
The steady state characteristics and dynamic performance obtained by a TOPCON laser controlled modular system, set up on an autograder performing an automatic leveling, and the ones supplied by an original test bench are at least comparable.
The original test bench designed and tested at INOE 2000-IHP from Bucharest allows the preliminar tuning of the laser controlled modular system for a given machine which will be turned into an automatic leveling equipment.
The test bench can be also used as a debugger for the leveling machines equipped with laser controlled modular systems as a fault detection tool. The special skilled staff can identify the component which does not provide anymore the required operational parameters: the laser transmitter, the laser receiver, the hydraulic block or the electronic block.
All the design parameters of the test bench were found by the aid of the numerical simulations performed with AMESIM. The facilities offered by this software for the engineering activities are turning this software into a real design tool. A lot of technical fields are developing high performance equipments, like speed governors for modern hydraulic turbines (Vasiliu et al., 2003), thrust vector actuators for aerospace control (Mare and Cregut, 2001), heavy load dynamic testing machines (Vasiliu and Vasiliu, 2004). Special tools as "activity index" for enhancing the synthesis process of the hybrid digital electro hydraulic control systems were developed by SOCIETE IMAGINE SA. The Real Time Simulation facilities of AMESIM widely extended the field of applications for this software (Vasiliu & Vasiliu, 2005).

8. References

Calinoiu, C., Vasiliu, N. & Vasiliu, D. (1998). Modeling, Simulation and experimental Identification of the Hydraulic Servomechanisms, *Technical Publishing House*, Bucharest, Romania, 222 p., ISBN 973-31-1315-8.

Lebrun, M. & Richards, C. (1997). How to create Good Models without Writing a Single Line of Code, *Proceedings of the Fifth Scandinavian International Conference on Fluid Power*, Linköping, Sweden.

LMS IMAGINE SA (2009). Advanced Modelling And Simulation Environment, Release 8.2.b., *User Manual*, Roanne, France.

The Math Works Inc. (2007). *Simulink R5,* Natick, MA, U.S.A.

Mare, J.C. & Cregut, S. (2001). Electro Hydraulic Force Generator for the Certification of a Thrust Vector Actuator, *Proceedings of the International Conference "Recent Advances in Aerospace Hydraulics"*, INSA Toulouse, France.

Popescu, T.C., Drumea, A. & Dutu, I. (2008). Numerical simulation and experimental identification of the laser controlled modular system purposefully created for equipping the terrace leveling installations, *Proceedings - Reliability and Life-time Prediction", (ISSE 2008)*, 7-11 May, 2008, Budapest, Hungary, ISBN: 978-963-06-4915-5; pp.336-341.

Popescu, T.C., Dutu, I., Vasiliu, C. & Mitroi, M. (2009). Adjustement of conformity parameters of PID-type regulators using simulation by AMESim, *Proceedings of the 7th International Industrial Simulation Conference 2009, ISC 2009*, June 1-3, 2009, Loughborough, United Kingdom, pp.269-274, Publication of EUROSIS-ETI.

Vasiliu, N., Călinoiu, C., Vasiliu, D., Ofrim, D. & Manea, F. (2003). Theoretical And Experimental Researches on a new Type of Digital Electro Hydraulic Speed Governor for Hydraulic Turbines, *1st International Conference on Computational Methods in Fluid Power Technology,* November, 26-28, 2003, Melbourne, Australia.

Vasiliu, N. & Vasiliu, D. (2004). Electro Hydraulic Servomechanisms with two Stages DDV for heavy Load Simulators controlled by ADWIN, *Proceedings of the International Conference "Recent Advances in Aerospace Hydraulics"*, INSA Toulouse, France.

Vasiliu, N. & Vasiliu, D. (2005). Fluid Power Systems, Vol.I., *Technical Publishing House,* Bucharest, Romania, ISBN 973-31-2249-1.

Part 5

Numerical Methods

21

A General Algorithm for Local Error Control in the RK*r*GL*m* Method

Justin S. C. Prentice
Department of Applied Mathematics, University of Johannesburg, Johannesburg, South Africa

1. Introduction

Simulation of physical systems often requires the solution of a system of ordinary differential equations, in the form of an initial-value problem. Usually, a Runge-Kutta method is used to solve such a system numerically. Recently, we examined how the computational efficiency of a Runge-Kutta method could be improved through the mechanism of the RK*r*GL*m* algorithm, in the context of global error control via reintegration (Prentice, 2009). The RK*r*GL*m* method for solving the *d*-dimensional system

$$\frac{d\overline{y}}{dx} = \overline{f}\left(x, \overline{y}\right) \qquad \overline{y}\left(x_0\right) = \overline{y}_0 \qquad a \le x \le b \tag{1.1}$$

is based on an explicit Runge-Kutta method of order r (RKr), and m-point Gauss-Legendre quadrature (GLm). The method has a global error of order $r+1$, which is the same order as the local order of the underlying RKr method, provided that r and m are chosen such that $r+1 \le 2m$ (Prentice, 2008). Of course, any method designed for solving IVPs must facilitate local error control. In this paper we describe an effective algorithm for controlling the local relative error in RK*r*GL*m*.

2. Terminology and relevant concepts

In this section we describe terminology and concepts relevant to the paper, including a brief description of the RK*r*GL*m* method. Note that, throughout this paper, *overbar*, as in $\overline{v}$, indicates an $d \times 1$ vector, and *caret*, as in $\widehat{M}$, denotes an $d \times d$ matrix.

2.1 Explicit Runge-Kutta methods

We denote an explicit RK method for solving (1.1) by

$$\overline{w}_{i+1} = \overline{w}_i + h_i \overline{F}\left(x_i, \overline{y}_i\right) \tag{2.1}$$

where $h_i \equiv x_{i+1} - x_i$ is a stepsize, $\overline{w}_i$ denotes the numerical approximation to $\overline{y}(x_i)$, and $\overline{F}\left(x, \overline{y}\right)$ is a function associated with the particular RK method (indeed, $\overline{F}\left(x, \overline{y}\right)$ could be regarded as the function that defines the method).

2.2 Local and global errors

We define the *global error* in any numerical solution at x_i by

$$\bar{\Delta}_i \equiv \bar{w}_i - \bar{y}_i \tag{2.2}$$

and, specifically, the *RK local error* at x_i by

$$\bar{\varepsilon}_i \equiv \left[\bar{y}_{i-1} + h_{i-1}\bar{F}\left(x_{i-1}, \bar{y}_{i-1}\right)\right] - \bar{y}_i \tag{2.3}$$

In the above, $\bar{y}_{i-1}$ and $\bar{y}_i$ are the true solutions at x_{i-1} and x_i, respectively. Note that the true value $\bar{y}_{i-1}$ is used in the bracketed term in (2.3).
Note also that for the derivative $\bar{y}' = \bar{f}\left(x, \bar{y}\right)$ we have

$$\bar{f}\left(x_i, \bar{w}_i\right) = f\left(x_i, \bar{y}_i + \bar{\Delta}_i\right) = \bar{f}\left(x_i, \bar{y}_i\right) + \widehat{f_y}\left(x_i, \bar{\vartheta}\right)_i \bar{\Delta}_i \tag{2.4}$$

In the above we use the symbol $\bar{\vartheta}_i$ in $\widehat{f_y}\left(x_i, \bar{\vartheta}_i\right)\bar{\Delta}_i$ simply to denote an appropriate set of constants such that $\widehat{f_y}\left(x_i, \bar{\vartheta}_i\right)\bar{\Delta}_i$ is the residual term in the first-order Taylor expansion of $\bar{f}\left(x_i, \bar{y}_i + \bar{\Delta}_i\right)$. Furthermore, $\widehat{f_y}$ is the Jacobian

$$\widehat{f_y} = \begin{bmatrix} \dfrac{\partial f_1}{\partial y_1} & \cdots & \dfrac{\partial f_1}{\partial y_d} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_d}{\partial y_1} & \cdots & \dfrac{\partial f_d}{\partial y_d} \end{bmatrix} \tag{2.5}$$

where $\{f_1, f_2, \ldots, f_d\}$ are the components of $\bar{f}$, and d is the dimension of the system (1.1). Clearly, a global error of $\bar{\Delta}_i$ in $\bar{w}_i$ implies an error of $O\left(\bar{\Delta}\right)_i$ in the derivative $\bar{f}\left(x_i, \bar{w}_i\right)$.

2.3 Gauss-Legendre quadrature

Gauss-Legendre quadrature on $[u, v]$ with *m* nodes is given by (Kincaid & Cheney, 2002)

$$\int_u^v \bar{f}\left(x, \bar{y}\right) dx = h \sum_{i=1}^{m} C_i \bar{f}\left(x_i, \bar{y}_i\right) + O\left(h^{2m+1}\right) \tag{2.6}$$

where the nodes x_i are the roots of the Legendre polynomial of degree m on $[u, v]$. Here, h is the average separation of the nodes on $[u, v]$, a notation we will adopt from now on, and the C_i are appropriate weights. The average node separation h on $[u, v]$ is defined by

$$h \equiv \frac{v - u}{m + 1}. \tag{2.7}$$

The nodes on $[-1, 1]$, denoted $\tilde{x}_i$, are mapped to corresponding nodes x_i on $[u, v]$ via

$$x_i = \frac{1}{2}\left[(v - u)\tilde{x}_i + u + v\right], \tag{2.8}$$

and the weights C_i are constants on any interval of integration. We have referred to the interval $[-1,1]$ above because the nodes $\tilde{x}_i$ on this interval are extensively tabulated.

2.4 The RK*r*GL*m* algorithm

We briefly describe the general RK*r*GL*m* algorithm on the interval $[a,b]$, with reference to Figure 1.

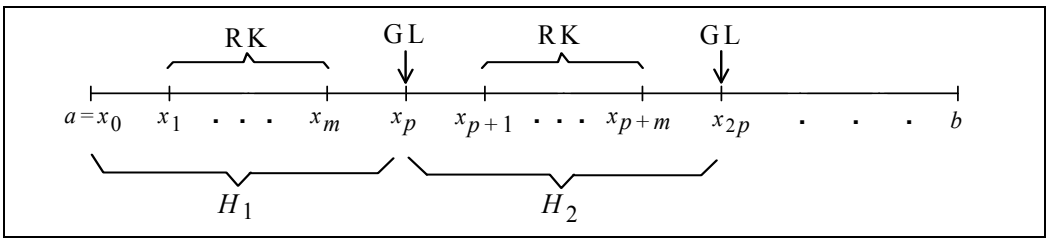


Fig. 1. Schematic depiction of the RK*r*GL*m* algorithm.

Subdivide $[a,b]$ into N subintervals H_j. At the RK nodes on H_j we use RK*r*:

$$\overline{w}_{i+1} = \overline{w}_i + h_i \overline{F}(x_i, \overline{w}_i) \qquad i \in \{(j-1)p, \ldots, (j-1)p+m-1\}. \tag{2.9}$$

At the GL nodes we use m-point GL quadrature:

$$\overline{w}_{(\mu+1)p} = \overline{w}_{\mu p} + h\sum_{i=1}^{m} C_i \overline{f}\left(x_{i+\mu p}, \overline{w}_{i+\mu p}\right). \tag{2.10}$$

where $\mu = 0,1,2,\ldots$. Note that $p \equiv m+1$.
The GL component is motivated by

$$\begin{aligned} \int_{x_{\mu p}}^{x_{(\mu+1)p}} \overline{f}(x,y)\,dx = \overline{y}_{(\mu+1)p} - \overline{y}_{\mu p} &\approx h\sum_{i=1}^{m} C_i \overline{f}\left(x_{i+\mu p}, \overline{y}_{i+\mu p}\right) \\ \Rightarrow \overline{y}_{(\mu+1)p} &\approx \overline{y}_{\mu p} + h\sum_{i=1}^{m} C_i \overline{f}\left(x_{i+\mu p}, \overline{y}_{i+\mu p}\right). \end{aligned} \tag{2.11}$$

The RK*r*GL*m* algorithm has been shown to be consistent, convergent and zero-stable (Prentice, 2008).

2.5 Local error at the GL nodes

The local error at the GL nodes is defined in a similar way to that for an RK method:

$$\begin{aligned} \int_{x_{\mu p}}^{x_{\mu p+m+1}=x_{(\mu+1)p}} \overline{f}(x,\overline{y})\,dx = \underbrace{\overline{y}_{\mu p+m+1}}_{(\mu+1)p} - \overline{y}_{\mu p} &= h\sum_{i=1}^{m} C_i \overline{f}\left(x_{i+\mu p}, \overline{y}_{i+\mu p}\right) + O\left(h^{2m+1}\right) \\ \Rightarrow \overline{\varepsilon}_{(\mu+1)p} \equiv \left[\overline{y}_{\mu p} + h\sum_{i=1}^{m} C_i \overline{f}\left(x_{i+\mu p}, \overline{y}_{i+\mu p}\right)\right] - \overline{y}_{(\mu+1)p} &= O\left(h^{2m+1}\right). \end{aligned} \tag{2.12}$$

We remind the reader that in RK*r*GL*m* we choose r and m such that $r+1 \le 2m$, which ensures that RK*r*GL*m* has a global error of $O\left(h^{r+1}\right)$ (Prentice, 2008).

2.6 Implementation of RK*r*GL*m*

There are a few points regarding the implementation of RK*r*GL*m* that need to be discussed:

a. If we merely sample the solutions at the GL nodes, treating the computations at the RK nodes as if they were the stages of an ordinary RK method, then RK*r*GL*m* would be reduced to an inefficient one-step method. This is not the intention behind the development of RK*r*GL*m*; rather, RK*r*GL*m* represents an attempt to improve the efficiency of any RK*r* method, simply by replacing the computation at every (m + 1)th node by a quadrature formula which does not require evaluation of any of the stages in the underlying RK*r* method.

b. Of course, it is clear from the above that on H_1 the RK nodes are required to be consistent with the nodes necessary for GL quadrature. If, however, the RK nodes are located differently (as would be required by a local error control mechanism, for example) then it is a simple matter to construct a Hermite interpolating polynomial of degree $2m+1$ (which has an error of order $2m+2$) using the solutions at the nodes $\{x_0,\ldots,x_m\}$. Then, assuming x_0 maps to -1 and x_m maps to the largest Legendre polynomial root x on $[-1,1]$, the position of the other nodes $\{x_1^*,\ldots,x_{m-1}^*\}$ suitable for GL quadrature may be determined, and the Hermite polynomial may be used to find approximate solutions of order $r+1$ at these nodes, thus facilitating the GL component of RK*r*GL*m*. A similar procedure is carried out on the next subinterval H_1, and so on. Indeed, we will see that the Hermite polynomial described here will play an important role in our error control process, and is described in more detail in the next subsection.

c. If the underlying RK*r* method possesses a continuous extension it would not be necessary to construct the Hermite polynomial described above. However, there is no guarantee that a continuous extension of appropriate order (at least $2m+1$) will be available, and it is generally true that determining a continuous extension for a RK method requires additional stages in the RK method, which would most likely compromise the gain in efficiency offered by RK*r*GL*m*. Note that the construction of the Hermite polynomial only requires one additional evaluation of $f(x,y)$, at x_m.

2.7 The Hermite interpolating polynomial

If the data $\{x_i, y_i, y_i' : i = 0,\ldots,m\}$ are available, then a polynomial $H_P(x)$, of degree at most $2m+1$, with the interpolatory properties

$$H_P(x_i) = y_i \qquad H_P'(x_i) = y_i' \tag{2.13}$$

for each i, may be constructed. If the nodes x_i are distinct, then $H_P(x)$ is unique. This approximating polynomial is known as the Hermite interpolating polynomial (Burden & Faires, 2001) and has an approximation error given by

$$y(x) - H_P(x) = \frac{y^{(2m+2)}\left(\xi(x)\right)}{(2m+2)!}\prod_{i=0}^{m}(x-x_i)^2 \tag{2.14}$$

where $x_0 < \xi(x) < x_m$. If h is the average separation of the nodes on $\left[x_0, x_m\right]$, it is possible to write $x - x_i = \sigma_i h$, where σ_i is a suitable constant, and hence

$$y(x) - H_P(x) = O(h^{2m+2}). \tag{2.15}$$

The algorithm for determining the coefficients of $H_P(x)$ is linear, as in

$$\mathbf{c} = \mathbf{A}^{-1}\mathbf{b} \tag{2.16}$$

where $\mathbf{c}$ is a vector of the coefficients of $H_P(x)$, $\mathbf{A}$ is the relevant interpolation matrix, and $\mathbf{b}$ is a vector containing y_i and y_i'. The details of these terms need not concern us here; rather, if an error $O(\Delta)$ exists in each of y_i and y_i', then an error of $O(\Delta)$ will exist in each component of $\mathbf{c}$. Moreover, since $H_P(x)$ is linear in its coefficients, then an error of $O(\Delta)$ will also exist in any computed value of $H_P(x)$. Consequently, we may write

$$y(x) - H_P(x) = O(h^{2m+2}) + O(\Delta) \tag{2.17}$$

where the $O(\Delta)$ term arises from errors in y_i and y_i'. We have assumed, of course, that the errors in y_i and y_i' are of the same order, which is the situation that we will encounter later.

3. Local error control in RK*r*GL*m*

3.1 The order of the tandem method

The idea behind the use of a *tandem* method is that it must be of sufficiently high order such that, relative to the approximate solution generated by RK*r*GL*m*, the tandem method yields a solution that may be assumed to be essentially exact. This solution is propagated in both RK*r*GL*m* and the tandem method itself, and the difference between the two solutions is taken as an estimate of the local error in RK*r*GL*m*. This amounts to so-called *local extrapolation* and is not dissimilar in spirit to error estimation techniques employed using Runge-Kutta embedded pairs (Hairer et al., 2000; Butcher, 2003). Generally speaking, though, the tandem method is not embedded.

To decide on an appropriate order for the tandem method we consider the local error at the GL nodes

$$\begin{aligned}\bar{\varepsilon}_{(\mu+1)p} &= \bar{y}_{\mu p} + h\sum_{i=1}^{m} C_i \bar{f}\left(x_{i+\mu p}, \bar{y}_{i+\mu p}\right) - \bar{y}_{(\mu+1)p} \\ &= \left(\bar{w}_{\mu p,t} - \bar{\Delta}_{\mu p,t}\right) + h\sum_{i=1}^{m} C_i \bar{f}\left(x_{i+\mu p}, \bar{w}_{i+\mu p,t} - \bar{\Delta}_{i+\mu p,t}\right) - \left(\bar{w}_{(\mu+1)p,t} - \bar{\Delta}_{(\mu+1)p,t}\right)\end{aligned} \tag{3.1}$$

where $\bar{w}_{(\bullet),t}$ is the solution from the tandem method at $x_{(\bullet)}$, and $\bar{\Delta}_{(\bullet),t}$ is the global error in $\bar{w}_{(\bullet),t}$. Expanding the term in the sum in a Taylor series gives

$$\begin{aligned}\bar{\varepsilon}_{(\mu+1)p} &= \bar{w}_{\mu p,t} + h\sum_{i=1}^{m} C_i \bar{f}\left(x_{i+\mu p}, \bar{w}_{i+\mu p,t}\right) - \bar{w}_{(\mu+1)p,t} \\ &\quad - \bar{\Delta}_{\mu p,t} + \bar{\Delta}_{(\mu+1)p,t} + h\sum_{i=1}^{m} C_i \widehat{f_y}(x_{i+\mu p}, \bar{\zeta}_{i+\mu p,t})\bar{\Delta}_{i+\mu p,t}\end{aligned} \tag{3.2}$$

and so

$$\begin{aligned}&\bar{w}_{\mu p,t}+h\sum_{i=1}^{m}C_i\bar{f}\left(x_{i+\mu p},\bar{w}_{i+\mu p,t}\right)-\bar{w}_{(\mu+1)p,t}\\&=\bar{\varepsilon}_{(\mu+1)p}+\left(\bar{\Delta}_{\mu p,t}-\bar{\Delta}_{(\mu+1)p,t}-h\sum_{i=1}^{m}C_i\widehat{f_y}(x_{i+\mu p},\bar{\zeta}_{i+\mu p,t})\bar{\Delta}_{i+\mu p,t}\right)\end{aligned}\tag{3.3}$$

where $\bar{\zeta}_{i+\mu p,t}$ is analogous to $\bar{\vartheta}_i$ in (2.4). The sum on the RHS of (3.3) is of higher order than $\bar{\Delta}_{\mu p,t}-\bar{\Delta}_{(\mu+1)p,t}$, because of the multiplication by h, and since we cannot expect, in general, that $\bar{\Delta}_{\mu p,t}-\bar{\Delta}_{(\mu+1)p,t}=0$, the term in parentheses must be $O\left(h^q\right)$, where q is the global order of the tandem method. Since $\bar{\varepsilon}_{(\mu+1)p}=O\left(h^{2m+1}\right)$ in the RKrGLm method, we require $q>2m+1$ in order for

$$\bar{w}_{\mu p,t}+h\sum_{i=1}^{m}C_i\bar{f}\left(x_{i+\mu p},\bar{w}_{i+\mu p,t}\right)-\bar{w}_{(\mu+1)p,t}\approx\bar{\varepsilon}_{(\mu+1)p}\tag{3.4}$$

to be a good (and asymptotically $\left(h\to 0\right)$ correct) estimate for the local error in RKrGLm. The first two terms on the LHS of (3.4) arise from RKrGLm with the tandem solution as input, while $\bar{w}_{(\mu+1)p,t}$ is the tandem solution at $x_{(\mu+1)p}$.
The implication, then, is that the tandem method must have a global order of at least $2m+2$, which implies $q>r+2$, since we already have $r+1=2m$ in RKrGLm. We acknowledge that our choice of q differs from conventional wisdom (which would choose $q>r+1$ so that the local order of the tandem method is one greater than the RK local order), but it is clear from (3.3) that the propagation of the tandem solution requires the global order of the tandem method to be greater than the order of $\bar{\varepsilon}_{(\mu+1)p}$. Of course, at the RK nodes the local order is $r+1$, so the tandem method with global order $q>r+2$ is more than suitable at these nodes.

3.2 The error control algorithm

We describe the error control algorithm on the first subinterval $H_1=[x_0(=a),x_p]$ (see Figure 1). The same procedure is then repeated on subsequent subintervals.
Solutions $\bar{w}_{1,r}$ and $\bar{w}_{1,q}$ are obtained at x_1 using RKr and RKq, respectively. We assume

$$\left|\bar{w}_{1,r}-\bar{y}_1\right|=\bar{L}_1h_0^{r+1}\approx\left|\bar{w}_{1,r}-\bar{w}_{1,q}\right|\tag{3.5}$$

where $h_0\equiv x_1-x_0$ and $\bar{L}_1$ is a vector of local error coefficients (we will discuss the choice of a value for h_0 later). The exponent of $r+1$ indicates the order of the local error in RKr. We find the maximum value of

$$\left|\frac{w_{1,r,i}-y_{1,i}}{y_{1,i}}\right|\approx\left|\frac{w_{1,r,i}-w_{1,q,i}}{w_{1,q,i}}\right|\qquad i=1,\ldots,d\tag{3.6}$$

where the index i refers to the components of the indicated vectors (so $w_{1,r,i}$ is the ith component of $\bar{w}_{1,r}$, etc). Call this maximum M_1 and say it occurs for $i = k$. Hence,

$$M_1 = \left| \frac{w_{1,r,k} - w_{1,q,k}}{w_{1,q,k}} \right| \tag{3.7}$$

is the largest relative error in the components of $\bar{w}_{1,r}$. Note that k may vary from node to node, but at any particular node we will always intend for k to denote the maximum value of (3.6). We now demand that

$$M_1 \le \delta_R \quad \Rightarrow \quad \left| w_{1,r,k} - w_{1,q,k} \right| \le \delta_R \left| w_{1,q,k} \right| \tag{3.8}$$

where δ_R is a user-defined *relative* tolerance. If this inequality is violated we find a new stepsize h_0^* such that

$$h_0^* = 0.9 \left(\frac{\delta_R \left| w_{1,q,k} \right|}{L_{1,k}} \right)^{\frac{1}{r+1}} \qquad \left(\Rightarrow \; L_{1,k} \left(h_0^* \right)^{r+1} < \delta_R \left| w_{1,q,k} \right| \right) \tag{3.9}$$

where $L_{1,k}$ is the kth component of $\bar{L}_1$, and then we find new solutions $\bar{w}_{1,r}$ and $\bar{w}_{1,q}$ using h_0^* ($\bar{L}_1$ is determined from (3.5)). The factor 0.9 in (3.9) is a *safety factor* allowing for the fact that $\bar{w}_{1,q}$ is not truly exact. To cater for the possibility that any component of $\bar{w}_{1,q}$ is close to zero we actually demand

$$\left| w_{1,r,k} - w_{1,q,k} \right| \le \max \left\{ \delta_A, \delta_R \left| w_{1,q,k} \right| \right\} \tag{3.10}$$

where δ_A is a user-defined *absolute* tolerance. We then set $h_1 = h_0^*$ and proceed to the node x_2, where the error control process is repeated, and similarly for x_3 up to x_m. The process of recalculating a solution using a new stepsize is known as a step rejection.

In the event that the condition in (3.10) is satisfied, we still calculate a new stepsize h_0^* (which would now be larger than h_0) and set $h_1 = h_0^*$, on the assumption that if h_0^* satisfies (3.10) at x_1, then it will do so at x_2 as well (however, we also place an upper limit on h_0^* of $2h_0$, although the choice of the factor two here is somewhat arbitrary). In the worst-case scenario we would find that h_1 is too large and a new, smaller value h_1^* must be used. The exception occurs when $\left| \bar{w}_{1,r,k} - \bar{w}_{1,q,k} \right| = \bar{0}$. In this case we simply set $h_1 = 2h_0$ and proceed to x_2.

The above is nothing more than well-known local relative error control in an explicit RK method using local extrapolation. It is at the GL node x_p that the algorithm deviates from the norm. A step-by-step description of the procedure at x_p follows:

1. Once error control at $\{x_1, x_2, \ldots, x_m\}$ has been effected (which necessarily defines the positions of $\{x_1, x_2, \ldots, x_m\}$ due to stepsize modifications that may have occurred), the location of x_p must be determined such that the local relative error at x_p is less than $\max \left\{ \delta_A, \delta_R \left| w_{p,q,k} \right| \right\}$, in which k has the meaning discussed earlier.

2. To this end, we utilize the map (2.8), demanding that $x_0 (= u)$ corresponds to -1 on the interval $[-1,1]$, and x_m corresponds to the largest root $\tilde{x}_m$ of the mth-degree Legendre polynomial in $[-1,1]$. This allows $x_p (= v)$ to be found, where x_p corresponds to 1 on $[-1,1]$, and so new nodes $\{x_1^*, x_2^*, \ldots, x_{m-1}^*\}$ can be determined such that $\{x_1^*, x_2^*, \ldots, x_{m-1}^*, x_m\}$ are consistent with the GL quadrature nodes on $[x_0, x_p]$.
3. We wish to perform GL quadrature, using the nodes $\{x_1^*, x_2^*, \ldots, x_{m-1}^*, x_m\}$, on $[x_0, x_p]$, but we do not have the approximate solutions $\{\bar{w}_{1,q}^*, \ldots, \bar{w}_{m-1,q}^*\}$ at $\{x_1^*, x_2^*, \ldots, x_{m-1}^*\}$.
4. Hence, we construct the Hermite interpolating polynomial $H_P(x)$ on $[x_0, x_m]$ using the original nodes $\{x_1, x_2, \ldots, x_m\}$ and the solutions that have been obtained at these nodes; of course, the derivative of $\bar{y}(x)$ at these nodes is given by $\bar{f}(x, \bar{y})$. Note that a Hermite polynomial must be constructed for each of the s components of the system, so if $d > 1$, $H_P(x)$ is actually a $d \times 1$ vector of Hermite polynomials.
5. We use the qth-order solutions that are available, so that we expect the approximation error in each $H_P(x)$ to be $O(h^q)$, as shown in (2.4) and (2.17).
6. The solutions $\{\bar{w}_{1,q}^*, \ldots, \bar{w}_{m-1,q}^*\}$ at $\{x_1^*, x_2^*, \ldots, x_{m-1}^*\}$ are then obtained from $\{H_P(x_1^*), \ldots, H_P(x_{m-1}^*)\}$.
7. GL quadrature then gives $\bar{w}_p$ with local error $O(h^{2m+1})$, as per (2.12).
8. The tandem method RKq is used to find $\bar{w}_{p,q}$, and $|\bar{w}_p - \bar{w}_{p,q}|$ is then used for error control:
   a. we know that the local error in $\bar{w}_p$ is $O(h^{2m+1})$, where h here is the average node separation on $[x_0, x_p]$;
   b. if the local error is too large then a new average node separation h^* is determined; using h^*, a new position for x_p, denoted x_p^*, is found from $x_p^* = x_0 + ph^*$;
   c. if $x_p^* > x_m$, we redefine the nodes $\{x_1^*, x_2^*, \ldots, x_{m-1}^*, x_m\}$, find qth-order solutions at these new nodes using $H_P(x)$, and then find solutions at x_p^* using GL quadrature and RKq;
   d. if $x_p^* \le x_m$, we reject the GL step since there is now no point in finding a solution at x_p^*.
9. After all this, the node x_p^* or x_m (if $x_p^* \le x_m$) defines the endpoint of the subinterval H_1; the stepsize h is set equal to the largest separation of the nodes on H_1, and the

entire error control procedure is implemented on the next subinterval H_2. Note also that it is the qth-order solution at the endpoint of H_1 that is propagated in the RK solution at the next node.

3.3 Initial stepsize

To find a stepsize h_0 to begin the calculation process, we assume that the local error coefficient $L_{1,k} = 1$ and then find h_0 from

$$h_0 = \left(\max\left\{\delta_A, \delta_R \left|\overline{y}_{0,k}\right|\right\}\right)^{\frac{1}{r+1}} \tag{3.11}$$

Solutions obtained with RKr and RKq using this stepsize then enable a new, possibly larger, h_0 to be determined, and it is this new h_0 that is used to find the solutions $\overline{w}_{1,r}$ and $\overline{w}_{1,q}$ at the node x_1.

3.4 Final node

We keep track of the nodes that evolve from the stepsize adjustments, until the end of the interval of integration b has been exceeded. We then backtrack to the node on $[a,b]$ closest to b (call it x_{f-1}), determine the stepsize $h_{f-1} \equiv b - x_{f-1}$, and then find $\overline{w}_{b,r}$ and $\overline{w}_{b,q}$, the numerical solutions at b using RKr and RKq, with h_{f-1}, x_{f-1} and $\overline{w}_{f-1,q}$ as input for both RKr and RKq. This completes the error control procedure.

4. Comments on embedded RK methods and continuous extensions

Our intention has been to develop an effective local error control algorithm for RK*r*GL*m*, and we believe that the above-mentioned algorithm achieves this objective. Moreover, the algorithm is general in the choice of RKr and RKq. These two methods could be entirely independent of each other, or they could constitute an embedded pair, as in RK(r,q). This latter choice would require fewer stage evaluations at each RK node, and so would be more efficient than if RKr and RKq were independent. Nevertheless, the use of an embedded pair is not necessary for the proper functioning of our error control algorithm.

The option of constructing $H_P(x)$ using the nodes $\{x_m = x_{p-1}, x_p, \ldots, x_{2p-1}\}$ for error control at x_{2p} (as opposed to using $\{x_p, \ldots, x_{2p-1}\}$) is worth considering. Such a polynomial, together with the Hermite polynomial constructed on $\{x_0, x_1, \ldots, x_m\}$, forms a piecewise continuous approximation to $\overline{y}(x)$ on $[x_0, x_{2p-1}]$. Of course, this process is repeated at the nodes $\{x_{2p}, x_{2p+1}, \ldots, x_{3p-1}\}$, and so on. In this way the Hermite polynomials, which must be constructed out of necessity for error control purposes, become a piecewise continuous (and smooth) extension of the approximate discrete solution. Such an extension is not constructed *a posteriori*; rather, it is constructed on each subinterval H_i as the RK*r*GL*m* algorithm proceeds, and so may be used for event trapping.

5. Numerical examples

We will use RK5GL3 to demonstrate the error control algorithm. In RK5GL3 we have $r = 5, m = 3$ so that the tandem method must be an eighth-order RK method, which we

denote RK8. The RK5 method in RK5GL3 is due to Fehlberg (Hairer et al., 2000), as is RK8 (Hairer et al., 2000; Butcher, 2003).
By way of example, we solve

$$y' = \frac{1}{1+x^2} - 2y^2 \tag{5.1}$$

on $[0,5]$ with $y(0) = 0$, and

$$y' = \frac{y}{4}\left(1 - \frac{y}{20}\right) \tag{5.2}$$

on $[0,30]$ with $y(0) = 1$. The first of these has a unimodal solution on the indicated interval, and we will refer to it as IVP1. The second problem is one of the test problems used by Hull et al (Hull et al., 1972), and we will refer to it as IVP2. These problems have solutions

$$\begin{aligned} &\text{IVP1:} \quad y(x) = \frac{x}{1+x^2} \\ &\text{IVP2:} \quad y(x) = \frac{20}{1+19e^{-x/4}} \end{aligned} \tag{5.3}$$

In Table 1 we show the results of implementing our local error control algorithm in solving both test problems. The absolute tolerance δ_A was always 10^{-10}, except for IVP1 with $\delta_R = 10^{-10}$, for which $\delta_A = 10^{-12}$ was used.

IVP1				
δ_R	10^{-4}	10^{-6}	10^{-8}	10^{-10}
RK step rejections	2	2	0	2
GL step rejections	2	5	10	19
nodes	12	20	37	79
RKGL subintervals	4	6	12	25

IVP2				
δ_R	10^{-4}	10^{-6}	10^{-8}	10^{-10}
RK step rejections	2	2	4	5
GL step rejections	2	3	5	9
nodes	10	19	39	87
RKGL subintervals	3	6	11	24

Table 1. Performance data for error control algorithm applied to IVP1 and IVP2.

In this table, *RK step rejections* is the number of times a smaller stepsize had to be determined at the RK nodes; *GL step rejections* is the number of times that $x_4^* \le x_3$, as described in the previous section; *nodes* is the total number of nodes used on the interval of integration, including the initial node x_0; and *RKGL subintervals* is the total number of subintervals H_i used on the interval of integration. It is clear that as δ_R is decreased so the number of nodes and RKGL subintervals increases (consistent with a decreasing stepsize), and so there is

more chance of step rejections. There are not many RK step rejections for either problem. When $\delta_R = 10^{-10}$ the GL step rejections for IVP1 are 19 out a possible 25 (almost 80%), but for IVP2 the GL step rejections number only about 38%). In both cases the GL step rejections arise as a result of relatively large local error coefficients at the GL nodes, which necessarily lead to relatively small values of h, the average node separation, so that the situation $x_4^* \le x_3$ is quite likely to occur.

Figures 2 and 3 show the RK5GL3 local error for IVP1 and IVP2. The curve labelled *tolerance* in each figure is $\delta_R |\overline{y}_i|$, which is the upper limit placed on the local error.

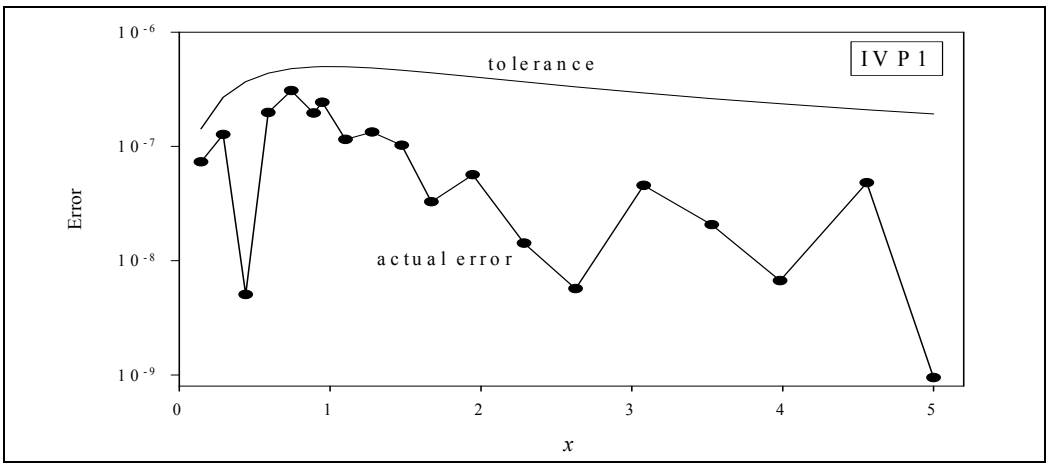


Fig. 2. RKGL local error for IVP1, with $\delta_R = 10^{-6}$.

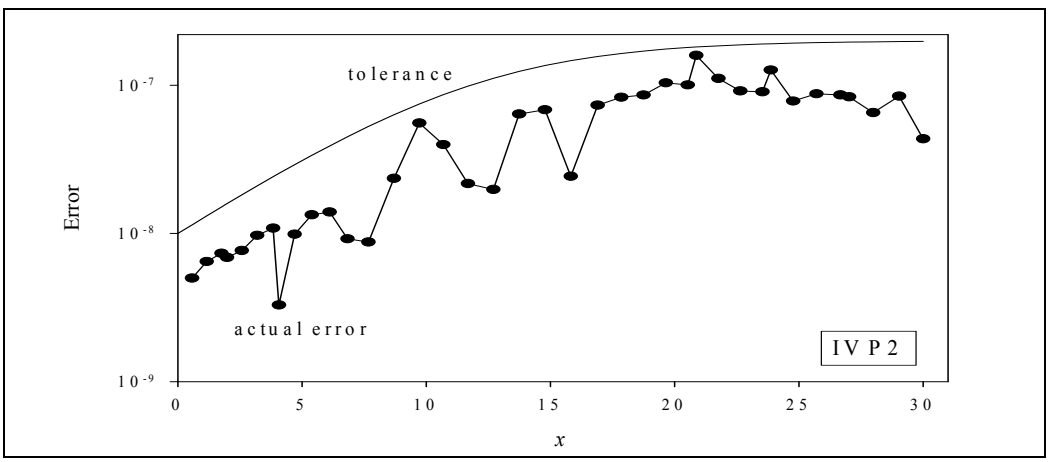


Fig. 3. RKGL local error for IVP2, with $\delta_R = 10^{-8}$.

In Figure 2 we have used $\delta_R = 10^{-6}$, and in Figure 3 we have used $\delta_R = 10^{-8}$. It is clear that in both cases the tolerance has been satisfied, and the error control algorithm has been successful. In Figure 4, for interest's sake, we show the stepsize variation as function of node index (#) for these two problems.

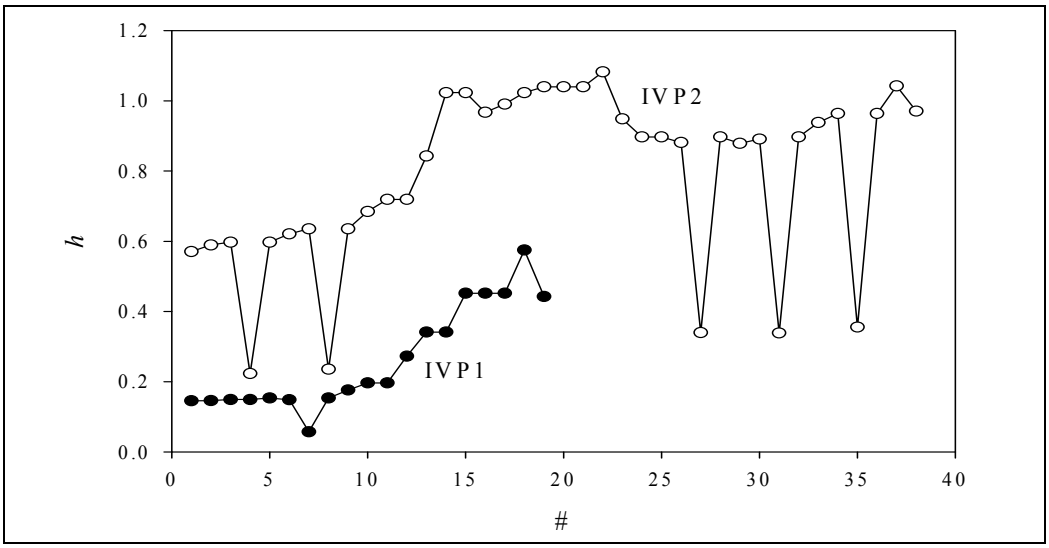


Fig. 4. Stepsize h vs node index (#) for IVP1 and IVP2.

To demonstrate error control in a system, we use RK5GL3 to solve

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= e^{2x}\sin x - 2y_1 + 2y_2 \\ y_1(0) &= -\frac{2}{5},\ \ y_2(0) = -\frac{3}{5} \end{aligned} \tag{5.4}$$

on $[0,3]$. The solution to this system, denoted SYS1, is

$$\begin{aligned} y_1 &= \frac{e^{2x}}{5}(\sin x - 2\cos x) \\ y_2 &= \frac{e^{2x}}{5}(4\sin x + 3\cos x) \end{aligned} \tag{5.5}$$

The performance table for RK5GL3 local error control applied to this problem is shown in Table 2.

SYS1				
δ_R	10^{-4}	10^{-6}	10^{-8}	10^{-10}
RK step rejections	3	5	6	9
GL step rejections	3	4	8	8
nodes	10	25	52	115
RKGL subintervals	3	7	15	31

Table 2. Performance data for error control algorithm applied to SYS1.

The performance is similar to that shown in Table 1. In all calculations reflected in Table 2, we have used $\delta_A = 10^{-12}$. The error in the components y_1 and y_2 of SYS1 is shown in Figures 5 and 6.

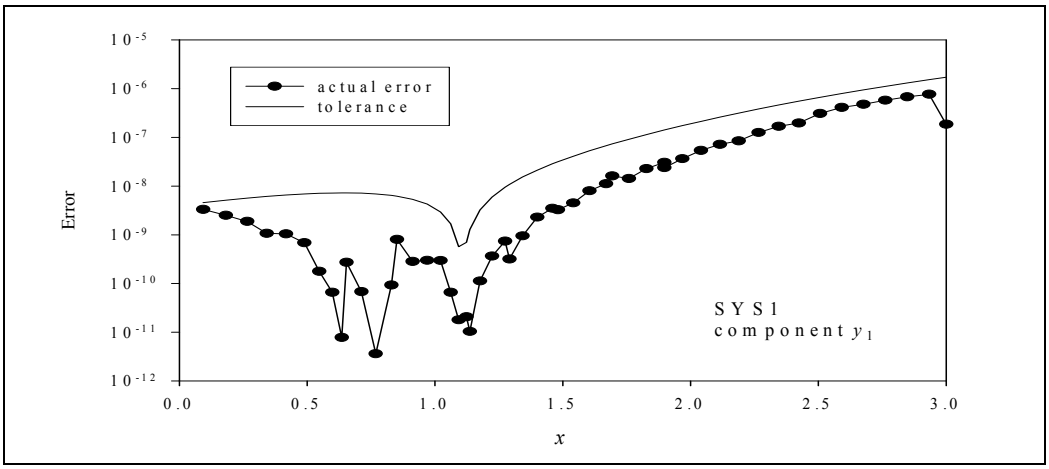


Fig. 5. Error in component y_1 of SYS1.

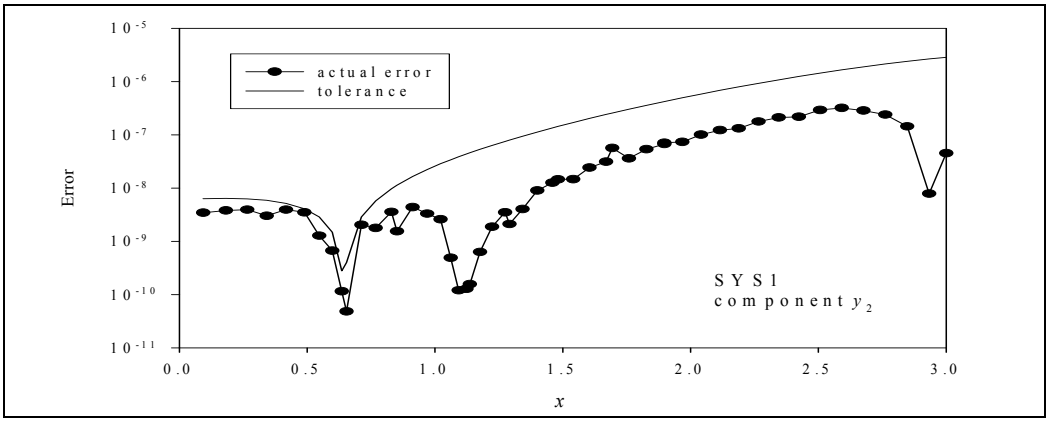


Fig. 6. Error in component y_2 of SYS1.

7. Conclusion and scope for further research

We have developed an effective algorithm for controlling the local relative error in RK*r*GL*m*, with $r+1 \le m$. The algorithm utilizes a tandem RK method of order $r+3$, at least. A few numerical examples have demonstrated the effectiveness of the error control procedure.

7.1 Further research

Although the algorithm is effective, it is somewhat inefficient, as evidenced by the large number of step rejections shown in the tables. Ways to improve efficiency might include :

a. The use of an embedded RK pair, such as DOPRI853 (Dormand & Prince, 1980), to reduce the total number of RK stage evaluations,
b. Using a high order RKGL method as the tandem method, since the RKGL methods were originally designed to improve RK efficiency,
c. Error control per subinterval H_j, rather than per node, which might require reintegration on each subinterval,

d. Optimal stepsize adjustment, so that stepsizes that are *smaller than necessary* are not used. Smaller stepsizes implies more nodes, which implies greater computational effort.

8. References

Burden, R.L. and Faires, J.D., (2001), *Numerical analysis, 7th ed.*, Brooks/Cole, 0-534-38216-9, Pacific Grove.

Butcher, J.C., (2003), *Numerical methods for ordinary differential equations*, Wiley, 0-471-96758-0, Chichester.

Dormand, J.R. and Prince, P.J., A family of embedded Runge-Kutta formulae, *Journal of Computational and Applied Mathematics*, 6 (1980) 19-26, 0377-0427.

Hairer, E., Norsett, S.P. and Wanner, G., (2000), *Solving ordinary differential equations I: Nonstiff problems*, Springer-Verlag, 3-540-56670-8, Berlin.

Hull, T.E., Enright, W.H., Fellen, B.M. and Sedgwick, A.E., Comparing numerical methods for ordinary differential equations, *SIAM Journal of Numerical Analysis*, 9, 4 (1972) 603-637, 0036-1429.

Kincaid, D. and Cheney, W., (2002), *Numerical Analysis: Mathematics of Scientific Computing, 3rd ed.*, Brooks/Cole, 0-534-38905-8, Pacific Grove.

Prentice, J.S.C., The RKGL method for the numerical solution of initial-value problems, *Journal of Computational and Applied Mathematics*, 213, 2 (2008) 477-487, 0377-0427.

Prentice, J.S.C.,Improving the efficiency of Runge-Kutta reintegration by means of the RKGL algorithm, (2009), In: *Advanced Technologies*, Kankesu Jayanthakumaran, (Ed.), 677-698, INTECH, 978-953-307-009-4, Vukovar.

22

Hybrid Type Method of Numerical Solution Integral Equations and its Applications

D.G.Arsenjev[1], V.M.Ivanov[1] and N.A. Berkovskiy[2]
[1]*Professor, St.Petersburg State Polytechnical University,*
[2]*assoc. prof., St.Petersburg State Polytechnical University,*
Russia

1. Introduction

The goal of current research is analysis of the effectiveness of application of semi-statistical method to the issues, which come up in computational and engineering practice.
The main advantages of this method are the possibility to optimize nodes on the domain of integration (which makes the work of calculator a lot easier), and also to control the accuracy of computations with the help of sample variance. Besides this, to improve the accuracy you can calculate an average solution by statistically independent estimations, acquired at a small number of integration nodes. A less attractive feature of this method is a low rate of convergence, which is relevant to all statistic methods.
The reason for this research has become a quite successful application of semi-statistical method to the test tasks [1, 2, 3]. The problem of plane lattice cascade flow with ideal incompressible fluid was chosen for simulation. With the help of semi-statistical method quite precise results have been achieved with the lattices, parameters of which were taken from engineering practice. These results were compared with the solutions from other computational methods.
Attempts to accelerate the rate of convergence brought to modernization of the method (deleting of spikes in average sum). As a result, in all the considered issues solutions with satisfactory precision were received, adaptive algorithm of lattice optimization was "putting" the nodes on the domain of integration in accordance with the theoretical considerations. However, in some cases the solution made by the semi-statistical method turned out to be longer, than when using deterministic methods, which is caused by imperfection of software implementation and also with the necessity to look for the new ways to accelerate rate of convergence for semi-statistical method, in particular, optimization mechanism.

2. Short scheme of semi-statical method

With semi-statistical method integral equations of the following kind can be solved:

$$\varphi(x) - \lambda \int_S K(x,y)\varphi(y)dy = f(x) \tag{1}$$

where S – smooth $(m-1)$-dimensional surface in R^m,
$x \in S$, $y \in S$, $\lambda \in R$,

K - kernel of equation, f - known function, φ- unknown function. This algorithm is described in detail in [1]. Let us shortly take a look at the scheme of its application in general case.

a. With the help of random number generator on the surface S. N - number of independent points $\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_N}$ (vectors) is created with a arbitrary probability density p(x) (random integration grid).
b. These points are placed one by one in (1), N equations of the kind given below are received:

$$\varphi(x_i) - \lambda \int_S K(x_i, y)\varphi(y)dy = f(x_i) \text{ , } (i=1,2,\ldots,N) \tag{2}$$

c. Integrals in (2) are substituted with the sums by the Monte-Carlo method [1, 2] and a system of linear algebraic equations appears

$$\varphi_i - \frac{\lambda}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{K(x_i, x_j)}{p(x_j)} \varphi_j = f(x_i) \tag{3}$$

Here $\{\varphi_i\}$ $(i=1,2,\ldots,N)$ vector of unknown variables of the system (3). Having solved (3), φ_i take for approximated value $\varphi(x_i)$ of the solution of integral equation (1) correspondingly. Approximated value φ(x) $\forall x \in S$ is defined by "retracing" with the following algorithm:

$$\varphi(x) \approx f(x) + \frac{\lambda}{N} \sum_{i=1}^{N} \frac{K(x, x_i)}{p(x_i)} \varphi_i \tag{4}$$

The bigger is N the more precisely integrals in (2) are approximated by the finite sum in (3), which means that we can suppose, that by incrementing value of N is possible to minimize calculating error of approximation of φ_i from (3) and φ(x) from (4) in a way that computation precision requires. As the number of thrown points is sometimes not enough to reach predefined precision (this number can't be enlarged infinitively as there is no possibility to solve to large equation systems), it is recommended to compute m times with N of thrown points, and then to average the results. This technique gives almost the same result if we would throw $N \times m$ points, because random points in different iterations are statistically independent.

d. You can get an estimated value of optimal density of integration nodes by formula [1] by means of approximated solution φ(x).

$$p_{onm}(y) = C(N-1) \frac{\sum_{j=1}^{N} \frac{\left(K(x_j, y)\varphi(y)\right)^2}{p(y)}}{\left| \sum_{j=1}^{N} \left\{ \frac{K(x_j, y)\varphi(y)}{p(x_j)} \sum_{\substack{i=1 \\ i \neq j}}^{N} K(x_j, x_i)\varphi(x_j) \right\} \right|} \tag{5}$$

Having generated the points with the density $p_{opt}(y)$, received from (5) by approximated values φ_i, we can get a more precise solution of the equation (1). After that with this equation and by means of (5) we can calculate again (more precise) value of optimal density. The process can be repeated till the density stops changing. This is the main sense of adaptive algorithm of choosing an optimal density.

3. Statement of the problem of blade cascade flow

A plane lattice with the increment t (Fig. 1) is given, on which from the infinity under the angle β_1 a potential flow of ideal fluid is leaking, coming out from a lattice under the angle β_2. The task is to find an absolute value of a normed speed of the flow on the edging of the profiles.

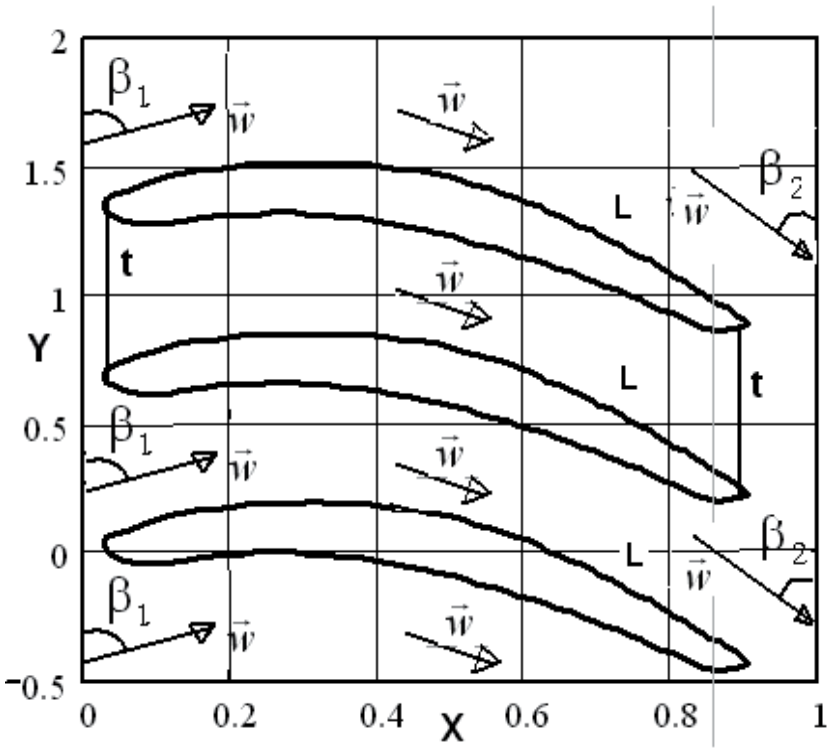


Fig. 1. Lattice of the profiles. $\vec{w}$ is a vector of the flow speed, t is an increment of the lattice, β_1 is and input angle of the flow, β_2 is and output angle of the flow, L is a contour of the blade profile

This task comes [5] to the solution of integral equation of the following kind:

$$w(s) + \oint_L \left(K(s,l) - \frac{1}{L} \right) \cdot w(l)\, dl = b(s), \tag{6}$$

where $w(s)$– normed speed of the flow;

$$K(s,l)=\frac{1}{t}\cdot\frac{\frac{\partial x}{\partial s}\cdot\sin\left(\frac{2\pi}{t}\cdot\left(y(s)-y(l)\right)\right)-\frac{\partial y}{\partial s}\sinh\left(\frac{2\pi}{t}\cdot\left(x(s)-x(l)\right)\right)}{\cos\left(\frac{2\pi}{t}\cdot\left(y(s)-y(l)\right)\right)-\cosh\left(\frac{2\pi}{t}\cdot\left(x(s)-x(l)\right)\right)};$$

$$b(s)=-2\cdot\frac{\partial x}{\partial s}-\frac{\partial y}{\partial s}\cdot\left(\cot(\beta_1)+\cot(\beta_2)\right)+\frac{t}{L}\quad .$$

Here s and l are values of the arch in different points of profile's edges, arches are counted from the middle of exiting border of the profile in the positive direction (counterclockwise); $x(l)$, $y(l)$ are the coordinates of the profile's point with the length of the arch l; **L** is a contour of the blade profile; ***L*** is the length of the contour of the blade profile.

The direction of the unit tangent vector $\left\{\frac{\partial x}{\partial s},\frac{\partial y}{\partial s}\right\}$ is chosen in a way that the tracking of the contour would be made counterclockwise. As opposed to [5], in this research front side of the lattice is orientated not along the abscises axis, but along the ordinate axis. Besides that, in [5] the speed is normed so that the flow expense of the fluid on the output would be unitary, and in this research the speed is normed so that the absolute value of speed vector on the outcome of the lattice would be unitary. This is achieved by multiplication of the speed, received after solving equations (1) and $\sin(\beta_2)$. Exactly the second norm rate setting is applied in the computational program of the Ural Polytechnic Institute (UPI), where the computations were made with the method of rectangles with the optimal setting of the integration nodes [2, 6] The solutions, received in this program, have been chosen for the comparison in this research.

4. Scheme of application of semi-statical method to the problem of blade cascade flow

4.1 Main formulas

In this task contour **L** acts as a surface S, and an integral equation (6) with an unknown function w(s) is solved. If by $w_k(s)$ we define an average solution after k iterations, and by $W_k(s)$ - value received on the integration with the number k after solving integral equation (1) on the N number of generated points, then we'll have

$$w_k(s)=\frac{1}{N}\cdot\sum_{m=1}^{k}W_k(s) \tag{7}$$

Selective standard deviation on the iteration with the number k is computated with the formula:

$$\delta_k(s)=\sqrt{\frac{1}{k^2}\cdot\sum_{l=1}^{k}D_l(s)}\ , \tag{8}$$

where D_l–selective dispersion in the end of one iteration with the number l;

$$D_l(s)=\frac{1}{N(N-1)}\sum_{m=1}^{N}\left(\frac{K(s,l_m)-\frac{1}{L}}{p(l_m)}\cdot W_k(l_m)+b(s)-W_k(s)\right)^2$$

Here $l_1, l_2, \ldots, l_N$ are random points on the segment [0, 2π], thrown on the iteration number k, N- a number of points in each iteration (in given below computational calculations is similar for all iterations), s-point of observation. Values of $w(l_m)$ are received as a result of approximated solution of integral equation (1) at the interation number k- of the method.
Computational practice has shown that deviation (calculating error) do not go behind the limits of standard deviation multiplied by three and, as usual, are within the boundaries of standard confidential (95%) interval.

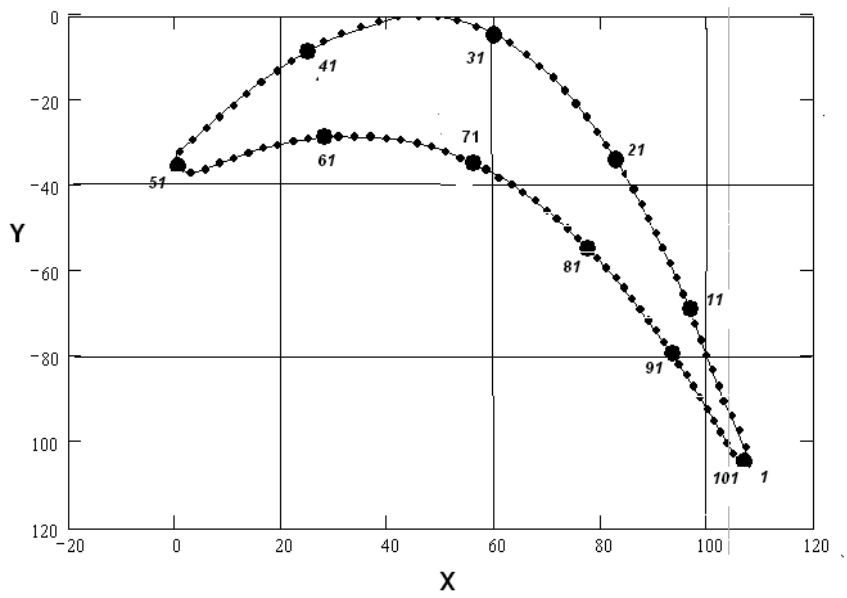


Fig. 2. Points, where the speed is calculated in computational examples – 50 equi-spaced points on the back and on the trough

Analytic definition of the blade contour

Integral equation (6) on the smooth contour **L** is of fredgolm type and has a unique solution [5]. Kernel of the equation (6) in case of two times differentiated contour can be considered continuous, as it has a removable singularity when $s=l$ [5]. In this research, however, contour is defined by a spline curve, first derivative of which has jumps in the finite set of points. This circumstance leads to the jumps of the kernel in the break points of the first derivative, which doesn't however influence the quality of computations. Besides that, the spline can always be approximated by a segment of Fourier row and the task can be solved on the infinitely derivated contour, as shown in [6].
Both approaches were tested, and the solutions made on a spline and on a segment of Fourier row were not considerably different. Semi-statical method was applied to the

equation (1) in accordance with the general scheme, which is described in detail in [3], without any additional preparation of regularization type. Values of the speed were calculated in 100 points of edging, 50 equal-spaced points on the trough and on the back correspondingly (Fig. 2). After that they were multiplied by $\sin(\beta_2)$, the result was compared to the solution, received by means of method of rectangles at the same contour. Both results were compared that to the one given by the UPI program. In the UPI program the contour is defined a little bit different, which causes insignificant divergences, which can be seen of the diagrams in the section of the results of computational modeling.

4.2 Computation algorithm and optimization.

The calculating was made iteratively. On each iteration a special number of random points on the segment [0, 2π] was generated with the density, which was calculated by the results of previous iterations (adaptive algorithm). On the first iterations points were generated with uniform density on the segment [0, 2π], which means approximately uniform distribution of the points on the contour of the blade, the results were defining more precisely by iterations, and approximated solution after iteration number i was considered to be arithmetic average of the solutions, received during previous iterations.

With the help of this approximated solution optimal density was calculated, using the method, described in [1]. Here algorithm is more economical, than described in [1], as it uses a more precise approximation to the right decision. As the computational practice has proved, on the strongly stretched contours on some iterations very strong spikes are possible, which are not smoothed by approximation even with the big number of iterations. However, it turned out that if the solution is very imprecise, than selective dispersions are also big in the check points, which are calculated during the work of the program.

We can introduce a constraint which will trace summands with a very big dispersion. In the current research the program is composed in a way that approximation is made not on all iterations, but only in those where relevant computational error, defined by the selective dispersion, is not bigger that 100 percent. Other solutions received on other iterations (usually not more than one percent from total number with the exclusion of the points close to edgings), are considered as spikes and are not included into the approximated finite sum. In case of much stretched working blade this improvement gives an undoubted advantage in the quality of computations.

With the help of semi-statistical method values of the speed were calculated in 150 points, distributed on the contour of the blade with an equal increment defined by the parameter u (which means practically equal increment on the arch length), and the values of the speed in checkpoints (which are distributed in the contour not evenly) were calculated with the help of interpolation. Selective dispersion is used as an index of precision of current approximation. It turned out that computer spends the most of time to calculating values of the kernel in the generated points, that's why the issue of decreasing number of generated points but saving precision of computation at the same time is important. In semi-statistical methods this can be achieved by optimization of the net of integration.

5. Results of computational modeling

To continue, let us introduce some denominations. On all the figures from 3 to 5 variable m stands for the number of point of observation, w_m is the speed in the point number m,

calculated by means of semi-statistical method, w_{1m} is the speed in point number m, calculated by means of method of rectangles, w_{2m} is the speed in point number m, calculated by means of UPI program, $|w_m - w_{1m}|$ is absolute deviation (calculation error) of calculation of the speed in point number m by means of semi-statistical method in comparison to the method of rectangles. Phrase "speed, calculated by means of semi-statistical method (4*400)" will mean that for calculation 4 iterations of semi-statistical method were made, with 400 points generated in every iteration.
Next (Fig. 3 – Fig. 5) the results of computational modeling are shown.

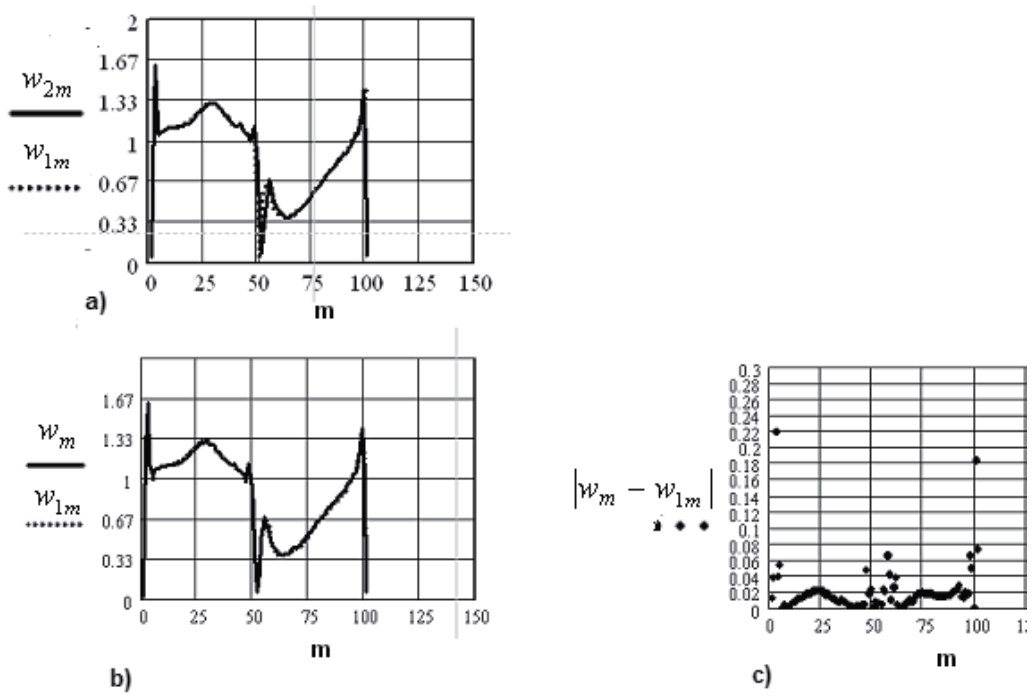


Fig. 3. Results of computational modeling on blade:
a) Speed graph, calculated by means of method of rectangles and speed graph calculated by means of UPI program
b) Speed graph, calculated by means of semi-statistical method (150*400) and speed graph, calculated by means of method of rectangles
c) Absolute deviation graph of calculation of the speed by means of semi- statistical method (150*400) in comparison to the method of rectangles

From given above examples (Fig.3) it is evident, that semi-statistical method commutated the speed with a good precision in all the points of contour, except for some points in the edgings, which are not important for practical issues.

6. Analysis of effectiveness of density adaptation

It was very interesting to investigate, how adaptive algorithm works when choosing optimal density. It appears that the points become denser on the edgings and on the back, which

means exactly the same places of profile, where the quality of computation is very bad during first iterations.
This is illustrated by the Fig. 4.

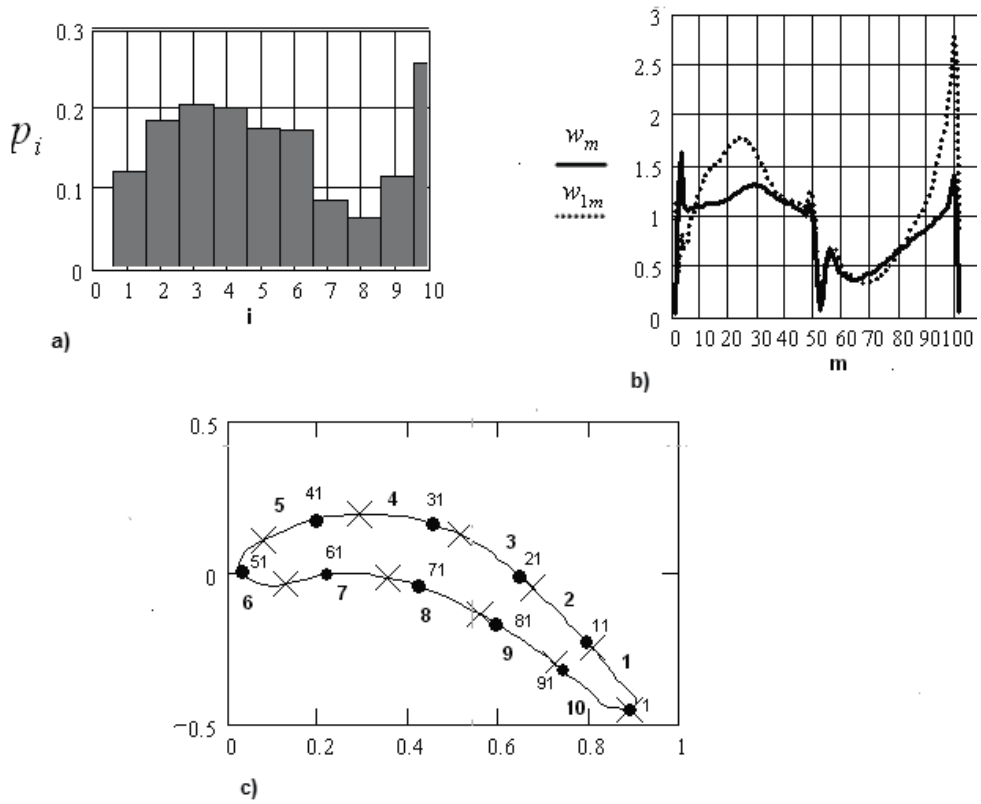
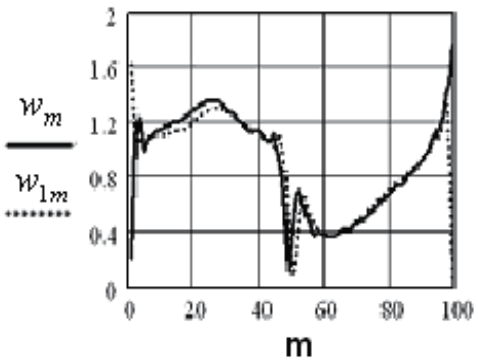


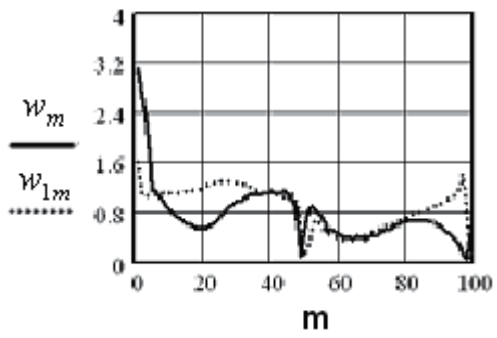
Fig. 4.

a) Histogram of optimal density after 2 iterations on the blade;
b) Speed graph calculated of the speed by means of semi-statistical method (2 iterations 400 points each) and speed graph calculated by means of method of rectangles.
c) With the symbol "×" borders of intervals from histogram on the Fig.3 a) are marked; numbers 1,2,…10 are the numbers of these intervals. Bold points are checkpoints (marked every 10 points starting with first); numbers 1,11,21,…92 – numbers of these points

On the Fig. 5 the results of computations on blade are shown, received after five iterations using adaptive algorithm for choice of optimal density and the results, received after five iterations with even distribution of generated points. It is easy to see that with the same number of generated points the results of adaptive algorithm are more precise.
From the Fig. 5 it is clear that using adaptive algorithm makes standard mean-square distance lower in shorter period, that with even distribution. It allows reducing the number of thrown points which is necessary to achieve predefined precision.



a)



b)

Fig. 5. Results of computational modeling on blade:
a) Speed graph, calculated by means of semi-statical method (5*400) and speed graph, with the use of adaptive algorithm, calculated by means of method of rectangles
b) Speed graph, calculated by means of semi-statical method (5*400) and speed graph, without the use of adaptive algorithm, calculated by means of method of rectangles

7. Conclusions

To sum up the results of computational modeling, following conclusions can be drawn:

a. By means of semi-statical method quite precise results can be achieved solving the problem of potential lattice cascade flow.
b. In accordance with theoretical computations adaptive algorithm works for optimization of nodes on the domain of integration. It fastens convergence, reducing selective dispersion.
c. However in strongly-stretched areas convergence rate is not very fast. The problem of fastening the rate of convergence, which is necessary to make semi-statical method successive in case of strongly-stretched areas and make it competitive to deterministic

methods in calculation speed, is still important. One of the ways to solve this problem is, evidently, improvement of the adaptive algorithm of optimization.

8. References

Arsenjev D.G., Ivanov V.M., Kul'chitsky O.Y. (1999). *Adaptive methods of computing mathematics and mechanics. Stochastic variant.* World Scientific Publishing Co., Singapore

Arsenjev D.G., Ivanov V.M., Korenevsky M.L. (2004). Semi-statistical and projection-statistical methods of numerical solving integral equations. *World Scientific and Engineering Academy and Society (WSEAS) Transactions on Circuits and Systems*, Issue 9, Volume 3: 1745-1749

Arsenjev D.G., Ivanov V.M., Berkovskiy N.A. (2004). Application of semi-statical method to inside Dirichle problem in three-dimensional space. *Scientific proceedings of SPbSPU*, St.Petersburg, № 4 (38):52–59

Isakov S.N., Tugushev N.U., Pirogova I.N. (1984). Report on scientific research. Development of computational software for calculating field of velocities and profile losses in lattices in compression in turbine buckets. *Proceedings of Ural polytechnical institute*, Svedrdlovsk

Jhukovskiy M.I. (1967). *Aerodynamic computing of flow in axial turbo-machines.* Mashgiz, Leningrad

Vochmyanin S.M., Roost E.G., Bogov I.A. (1997). *Computing cooling systems for gas turbine blades. Bundled software GOLD*. International academy of high school, Petersburg

Part 6

Safety Simulation

23

Advanced Numerical Simulation for the Safety Demonstration of Nuclear Power Plants

G.B. Bruna, J.-C. Micaelli, J. Couturier, F. Barré and J.P. Van Dorsselaere
Institut de Radioprotection et de Sûreté Nucléaire,
31 avenue de la Division Leclerc, 92260 Fontenay-aux-Roses
France

1. Introduction

Existing commercial nuclear power plants (NPPs) have obtained excellent and outstanding performance records over the past decade. Nevertheless, even though the high safety level already achieved could be maintained without investing new exhaustive research efforts, anticipation of further tighter requirements for even higher standard levels should be made, which implies preparedness for new research. Accordingly, in the near and intermediate future, research will conceivably focus on new emerging trends as a result of further desire to reduce the current uncertainties for better economics and improved safety of the current reactors and requirements of the new reactor designs.

As it has been usual in the past, the research will continue serving the short-term needs of the end-users (regulatory bodies, utilities and vendors) which mainly focus on both emerging and pending issues, but it will also contribute to addressing the long-term safety needs or the questions arising from the changes in plant designs and operating modes, and to preparing the emergence of new concepts. The sensibility of the stakeholders for a continuous enhancement of safety, mainly when dealing with the advanced and innovative concepts, will entail the development of reactor concepts able to intrinsically prevent severe accidents from occurring, and, should that not be possible, reduce either their probability or the level of expected consequences on the environment and the populations. That should be done in first priority by design, and not necessarily by improvements or the addition of safety systems.

Such anticipatory research will involve new generation simulation tools and innovative experimental programs, to be carried out both in the research facilities currently in operation throughout the world and in new dedicated mock-ups supported by suitable laboratory infrastructures. Enhanced or complementary data banks to be generated and further investigations on human and organizational factors will be the primary research activities, from which the end users will definitely profit.

In addition, significant efforts should be devoted to get the maximum benefit from the computation tools already available and start preparing their improvements as well by taking advantage from the development and availability of new computation techniques, such as advanced numerical simulation.

Their applicability should be extended to all types of current and future water cooled reactors and validated under the conditions of new designs. Such an "extrapolation" of the already gathered knowledge in the field of Light Water Reactors (LWRs) would maximize benefit from the work already done and could save some major efforts in the future.

2. Numerical simulation in the current nuclear safety context

In a context of a worldwide renaissance of nuclear energy, the most important pending milestones for the Generation II reactors are the periodic safety review (every ten years in France), which include safety reassessments, as well as the demand for long-term operation of the plants - far beyond their original design lifetime -. Additionally, the safety assessment for Generation III and III+ reactors under construction must be carried out and the safety demonstration for future highly innovative Generation IV (GEN IV) reactors accurately prepared.

On the other hand, over the past decade, considerable progress has been made in the domain of numerical simulation in many fields of endeavor.

In this challenging context, the following two main questions are raised:

- Could the safety demonstration of current and future power plants benefit from the progress currently made in numerical simulation?
- Does the safety demonstration of GEN IV concepts require a breakthrough in terms of numerical simulation?

This chapter is intended to address both questions and provide with preliminary elements of answer. In its first part, through some selected examples, it illustrates the development perspectives for the computation tools that are currently adopted in the safety demonstration of nuclear power plants, and wonders about the future contribution to these tools of the progress made in advanced numerical simulation. In its second part, for a selected sample of GEN IV concepts, it investigates the directions the modeling efforts (including advanced simulation ones) could and/or should be orientated towards.

At least two ways for progress (which are not mutually exclusive) are identified in the development of computation tools already adopted or to be adopted for current reactor concept design and safety studies:

- The first one relies on a progressive sophistication of the physical models, the codes adopted for Loss Of Coolant Accidents (LOCA) transient studies providing a wide field of application.
- The second one holds on advanced detailed modeling. It includes:
  - The investigation of phenomena at a physical scale significantly smaller than for the current generation of safety codes. It may contribute, through the so-called multi-scale approaches, to improving the macroscopic models (as it is presently the case for the fuel), and/or, whenever possible, to replacing them. A pertinent example in the field of severe accidents is the current use of Computational Fluid Dynamics (CFD) codes to investigate the risk of hydrogen explosion in the containment.
  - The coupling of different physical fields. Pertinent examples can be found in the domain of reactivity accidents, including dilution accidents: for these transients, such as un-borated water injection at shutdown, more accurate methodologies are now under development, they allow coupling different fields contributing to the power excursion (neutronics, fuel thermal-mechanical and thermal-hydraulics).

As far as the GEN IV concepts are concerned, today in France only 3 out of the 6 concepts proposed by the GEN IV International Forum (GIF) are currently considered:

- The Sodium Fast Reactor (SFR) that benefits from significant industrial and operating experience in several countries, including France;
- The Gas Fast Reactor (GFR) that possesses a very high potential in terms of uranium sparing, incineration, transmutation and heat production;
- The High or Very High Temperature Reactor (HTR/VHTR) that is the most likely concept to be inherently safe and multi-use and benefits from a first industrial experience in several countries.

Each of these concepts, according to its physical features and operating mode, engenders specific needs in terms of development and assessment of computation tools. Nevertheless, several major trends can be mentioned as relevant to the safety demonstration and widely independent from the design. At the present and first stage of IRSN's investigation, 5 main issues have been pointed out: the consistence and robustness of neutronics design, the demonstration of the actual capacity to passively and safely evacuate the residual power, the fuel integrity, the quantification of activated fission products that might be released to the environment in case of accidental situations, the inquiry upon the significant reduction of a likelihood of severe core damage, particularly the prevention of the "design basis" conditions from degenerating into severe accidents.

All of them could benefit from the current progress in advanced simulation. The chapter accurately investigates the potential contribution of progress in numerical simulation, and more specifically the advanced one, to the above-mentioned safety issues.

3. Current practice of advanced numerical simulation in nuclear safety

Before addressing the numerical simulation for the safety demonstration of GEN IV concepts, it is worthwhile presenting a quick overview of the present status concerning the use of advanced numerical simulation techniques in current nuclear safety analysis. This status has already been discussed and elaborated in specific seminars and workshops, e.g. the meeting organized by OECD and IAEA (OECD IAEA, 2002) for CFD, as well as in a previous IRSN's paper (Livolant, et al. 2003).

In the following, some LWRs safety related topics are addressed such as: Primary circuit thermal-hydraulics and LOCA, Fuel behavior in Design Basis Accident (DBA), Coupled phenomena in DBA, Severe accidents (SA), and Use of CFD codes in other accidents.

3.1 Primary circuit thermal-hydraulics and Loss Of Coolant Accidents (LOCA)

A key safety problem in LWRs is guaranteeing the coolability of the fuel in any normal operation, incidental and accidental condition, including the worst case of a pipe rupture. The development of codes able to treat the problem with some realism started in the early 70s. At that time, the main challenge was calculating the behavior of a steam-water flow in a hot pressurized circuit, with a breach into the containment building.

Today, various code systems are internationally available. Their physical models are based on experimentally-supported reasonable assumptions on the steam and water flows as well as their mutual interactions. The circuits are represented assembling together 1D pipe elements, 0D volumes, and, whenever possible, 3D components. In the past, intensive experimental programs to validate these codes have been carried out either on the system loops or on components mock-ups. As a consequence, a sufficient and convenient

confidence level exists in their results, at least when they are used within their validation domain and by skilled users.
The calculation results significantly improve the safety analysis and the probabilistic risk analysis. The existing codes are able to offer a satisfactory answer for the reactors in operation and even for the next generation of evolutionary water reactors (GEN-III). The lasting requirements for improvement mainly concern their robustness, reliability and user friendliness.
However, the confidence in the results of these codes widely relies on their experimental validation. Extrapolation to situations out of the validation domain may provide doubtful and sometimes even erroneous results. So, for both design and safety reasons, in presence of significant design and operation changes, it would be worth improving the existing modeling. An international consensus exists on the interest to keep maintaining an R&D activity aimed at achieving that objective.
In the medium term (5 to 10 years), the two-fluid models are expected to improve with extension to fields like droplets, and incorporation of transport equations for the interfacial area, and the 3D modeling should be extended as far as possible. This strategy is likely to sustain a process of progressive improvement, without any significant breakthrough.
Meanwhile, the increasing computer efficiency should allow using refined meshing and capturing smaller scale phenomena, provided that convenient models are made available. In this regard, it is worth recalling that the study of non-azimuthal cold shocks on reactor vessel of the first generation French Pressurized Water Reactors (PWR) (900 MWe) has been performed by the French Utility (EDF - *Electricité de France*) with CFD codes. Nevertheless, conventional system and component codes are likely to remain the basic tools for long, while benefiting from the development of the more refined approaches derived from CFD codes and Direct Numerical Simulation (DNS).
CFD codes will allow zooming on specific zones of a circuit or may be used as a powerful investigation tool to derive new closure relationships for more macroscopic approaches, thereby reducing the need for expensive experimental programs. Coupling between CFD and system codes may also be an efficient way to improve the description of small-scale phenomena while maintaining computer costs and time consumption at reasonably low levels.
Once the underway developments are available, the DNS codes will be adopted to search for a better understanding of small scale physical processes and derive new and more accurate models for averaged approaches.
In conclusion, the strategy for preparing the next generation of thermal-hydraulic tools consists in improving the capabilities of system and component codes by developing new models while extending CFD codes capabilities to all flow regimes and improving DNS techniques. Nevertheless, the concern for the uncertainties in CFD simulations is still to be addressed.

3.2 Fuel behavior in DBA

A major safety concern in LWRs is the possible failure of core fuel rods during transients, such as a LOCA or a RIA (Reactivity Initiated Accident, which can be initiated for example by an uncontrolled control rod withdrawal). Such failures can modify the core geometry and reduce its coolability; they can also engender the ejection of fuel fragments (and consequently radioactivity) in the reactor primary circuit. During the 60s and in the early

70s, several experimental programs were carried out, which provided information about fuel rods behavior. The results were used to develop and assess RIA and LOCA fuel codes.
At that time, the fuel was pure UOX (Uranium Oxide) and the burn-up was limited to 40GWd/kg; data for low burn-up had been included in data bases for code assessment, and it was believed that some extrapolation in burn-up was acceptable. By the mid-1980s, however, significant changes in the pellet microstructure and clad mechanical properties were observed in experiments carried out with fuel at higher burn-up and MOX (Mixed Oxide, i.e. containing both Uranium Oxide and Plutonium Oxide).
Those observations provided evidence that the fuel thermal-mechanical behavior is strongly dependent on the fuel type (UOX, MOX, etc.) and the cladding material, and that extrapolation was not always appropriate. Thus, a large number of experimental and analytical programs were initiated to check the fuel behavior and model the effects of the higher burn-up of fuel elements proposed by fuel designers, mainly under RIA and LOCA conditions.
Fuel codes for RIA analysis include models, correlations, and properties for cladding plastic stress-strain behavior at high temperatures, effects of annealing, behavior of oxides and hydrides during temperature ramps, phase changes, and large cladding deformations such as ballooning. The mechanical description of cladding should preferentially be 2-dimensional, but models of lower dimension are used as well; moreover, it generally includes a failure model. These codes also include fuel pellets thermal-mechanical models that may interact with fission gas models.
The mechanical models of pellets are generally mono-dimensional. Special care is to be paid to the modeling of the so-called pellet RIM-zone (i.e. the very external boundary of it where most of nuclear interactions currently occur) and the MOX due to its heterogeneous nature (the MOX grains – of quite large size - are dispersed in a UOX very thin matrix).
Fuel codes for LOCA analysis usually adopt built-in heat transfer correlations (cladding to coolant), a constant or dynamic gap conductance model, and average values for thermal conductivity and heat capacity. As regards clad thermal-mechanical aspects, these codes typically describe ballooning and include burst and oxidation models. Although simpler in the practice, the LOCA fuel models take into account high burn-up effects and thermal-mechanical characteristics of different types of fuel elements. New specific developments are underway to treat fuel relocation, an important phenomenon recently highlighted in the framework of the OECD-Halden program (OECD/NEA, 2003).
DRACCAR is currently developed at IRSN for the simulation of the thermal-mechanical behavior of a rod bundle under LOCA conditions, with a 3D multi-rod description (Figure 1). The objectives are to simulate mechanical and thermal interactions between rods, to evaluate the blockage ratio, as well as the structure embrittlement and the coolability of the fuel assembly. The reflooding phase of a fuel rod assembly during a LOCA transient can be calculated when DRACCAR is coupled with a suitable thermal-hydraulics code. The models are applicable to any kind of fuel (UO_2, MOX …), cladding (Zircaloy 4, Zirlo, M5 …), core loading and management (burn-up …) and types of water-cooled reactor (PWR, Boiling Water Reactor or BWR, …). It is also applicable to fuel handling or spent fuel pool draining accidents. A version for GEN IV SFR is planned. The flexibility of the DRACCAR code allows to model from one single rod to a fuel assembly. Each structure is in mechanical and thermal interaction with others, including contacts between fuel rods and eventually with guide tubes. Each rod has a 3D description and is coupled with a sub-channel thermal-hydraulics. The code uses 3D non structured meshing to describe the fuel assembly.

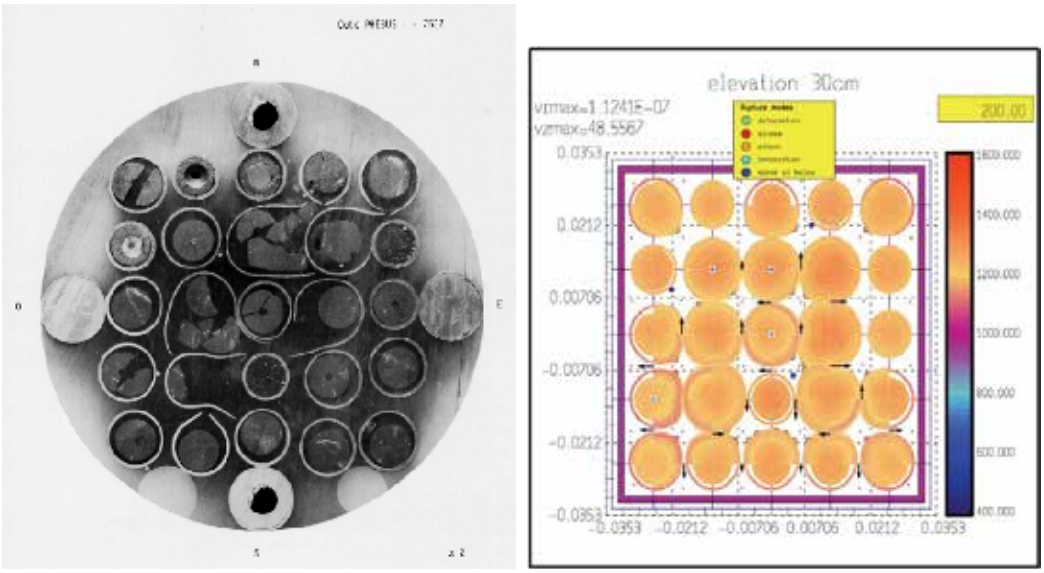


Fig. 1. Bundle deformation obtained during Phebus LOCA tests (run 215), 5 x 5 rod bundle; experimental results and DRACCAR simulation

Even if a limited number of model improvements are still judged necessary in the fuel codes, it is widely agreed that these developments could be achieved without any major breakthrough; however, it is to be mentioned that in order to improve the physical basis of models and consequently to give some confidence in extrapolations (beyond the domain covered by experimental results) the fuel models are more and more often backed up by the above-mentioned multi-scale approach.

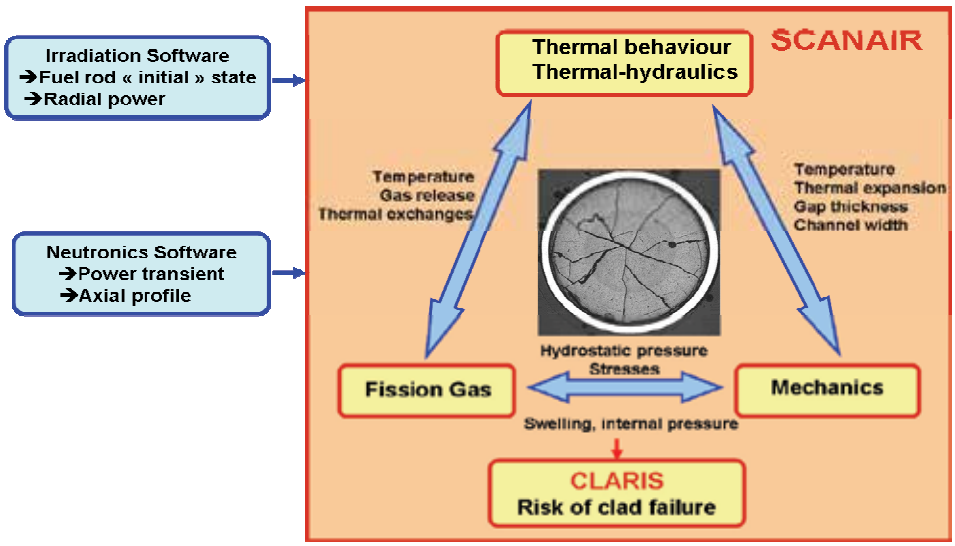


Fig. 2. The SCANAIR computation principle

A typical application of such an approach can be found in the work carried out at IRSN to improve the modeling of the Zircaloy clad behavior. This entails modeling cladding behavior on a micro-scale that represents the structures composing the cladding. In this case, the characteristic size is set by the thickness of the zirconium hydride disks (form in which the hydrogen diffused in the cladding precipitates). As the structure is subdivided into elementary units, behavior laws have to be established for each one of them. Homogenization methods were used to determine the current volume behavior of the material.

These improvements were undertaken to develop an anisotropic elastoplastic behavior model for hydride Zircaloy that may be used at macro-scale in current RIA fuel codes such as SCANAIR developed by IRSN in the framework of a collaboration with EDF, and globally assessed on CABRI REP-Na (Papin et al., 2007) and NSRR (Suzuki et al., 2006) in-pile or integral experiments.

SCANAIR is a thermo-mechanical code simulating a fuel rod surrounded by coolant that undergoes an RIA (Figure 2). The SCANAIR code couples three modules: the first one calculates fission gas migration and release into the rod gap, the second one deals with mechanics (it calculates the stresses and strains in the fuel and in the cladding) and the third one evaluates the fuel, cladding and coolant temperatures.

The use of multi-scale approaches should increase the confidence in the extrapolations from experimental conditions to reactor ones. It should also contribute to optimizing the definition of the experimental programs and decreasing their global cost; nevertheless, the necessity of code assessment against so-called "integral" experiments (i.e. experiments involving all the major phenomena that could occur in reactor conditions) will remain to verify the consistency of the different models (in particular models that have been independently derived by multi-scale approaches) and check that there is no important omission.

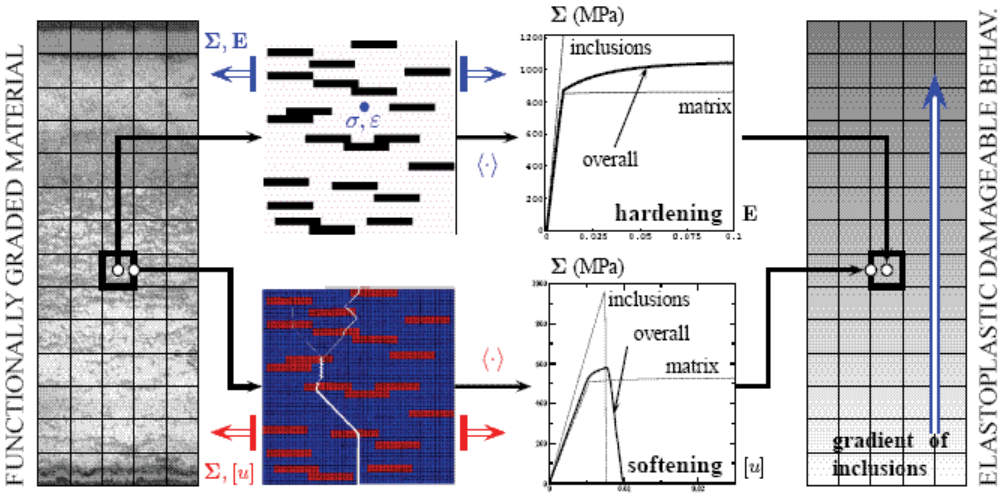


Fig. 3. Principle of the multiscale cohesive-volumetric approach for the study of the overall elastoplastic and damageable behavior of a functionally graded material

In the recent years, a new approach has been developed to predict the ductile fracture of heterogeneous materials during transient loadings. This approach is based on the so-called cohesive volumetric finite element (CVFE) method in the periodic homogenization framework (Perales et al., 2008). The coupling of this numerical approach to some analytical homogenization models allows predicting the behavior of heterogeneous materials from elasticity to ductile damage up to failure.
The framework of this coupling has been applied to a material from the nuclear industry: the highly irradiated Zircaloy cladding. This application illustrates a coupled approach where the overall hardening behavior of a composite material (as elastoplasticity) is incorporated into the bulk behavior and the overall softening behavior (as damage and fracture) is incorporated into some cohesive zone models.
The highly irradiated Zircaloy cladding is a functionally graded material composed of a metal matrix and aligned brittle hydride inclusions (Figure 3). The overall elastoplastic and damageable behavior of this material is obtained using the CVFE method where both the mean volumetric and cohesive properties arise from homogenization techniques at the micro-scale. The volumetric hardening behavior is obtained adopting a homogenization model based on a variational approach, and the cohesive softening behavior comes from a periodic CVFE modeling (Perales et al., 2006).

3.3 Coupled phenomena in DBA

Compliance with safety criteria in DBA and, more generally, in any operation, incidental and accidental circumstance of the reactor life requires the development of neutronics, fuel thermal-mechanical and thermal-hydraulics models. In principle, these three fields should be accounted for simultaneously because:

- The neutron cross-sections depend on the fuel temperature and the moderator density;
- The fuel temperature depends on the fuel element geometry, the neutronics power and the thermal exchange with the moderator fluid;
- The thermal-hydraulics depends on the fuel element geometry, the "source term" corresponding to the power released by convection and by γ radiation.

Up to now, due to the heaviness and complexity of computations, the methods adopted in the safety analysis have assumed these three fields as more or less decoupled. The major disadvantage of this assumption is the impossibility to accurately compute the pin-wise power distribution of the core. Thus, power peaking factors are adopted for design and safety analysis. Whereas they are evaluated in steady-state conditions, they are used for transient studies adding some corrections to ensure conservatism.
Incorporating full three-dimensional (3D) models of the core in the system transient codes enables the interactions between the core behavior and the plant dynamics to be accounted for in a more consistent way. Recent progress in computer technology has been achieved in the development of coupled thermal-hydraulics, fuel thermo-mechanical behavior, neutron kinetics and system codes.
Developments of several multi-physics code systems are currently underway, among which the NURESIM platform being developed in the frame of the 6th Framework R&D program of the European Commission and the HEMERA (Highly Evolutionary Methods for Extensive Reactor Analyses) coupled chain, developed jointly by IRSN and CEA (Figure 4). The HEMERA chain (Bruna et al., 2007) features are intended to allow performing more accurate calculations for the safety assessment of the thermal nuclear reactors in operation, in association with uncertainty and sensitivity studies and penalization techniques.

The HEMERA computation chain is a fully coupled 3D code system developed jointly by IRSN and CEA. It comprises the CRONOS neutronics code, the FLICA thermal-hydraulics code and the CATHARE system code. The ISAS supervisor manages the coupling. The nuclear data (neutron cross-sections) are provided to HEMERA by the APOLLO-2 code. HEMERA allows performing coupled (neutronics/thermal-hydraulics) calculations.

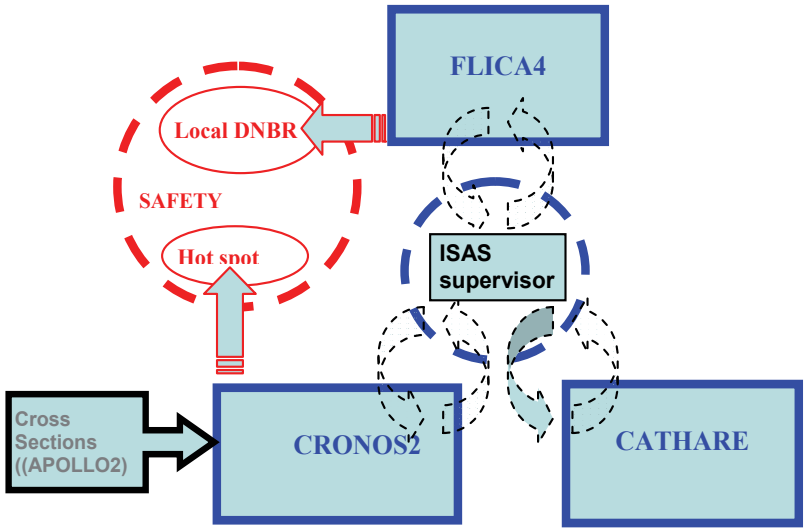


Fig. 4. The HEMERA computation chain

Accident analyses should demonstrate compliance with safety criteria. As far as the simulation of transients is concerned, the traditional French approach compels adopting the most penalizing initiators, so that neutronics, thermal and thermal-hydraulics calculations have to be either externally or internally coupled. HEMERA provides this coupling internally through the multi-level and multi-dimensional models which have been implemented to account for neutronics, core thermal-hydraulics, fuel thermal analysis and system thermal-hydraulics phenomena with best estimate and/or conservative assumptions (Clergeau et al. 2010).

3.4 Severe accidents

Historically, for a long time the LOCA has been considered as the maximum credible accident in LWRs. Accordingly, their main safety design features have been defined to prevent it or, at least, to limit its consequences, through keeping the core geometry coolable as long as possible, and strictly limiting the fission products release to the environment.

However, since the 70s, and mainly as a consequence of the TMI2 accident, it was internationally agreed that it is necessary to account for accidental situations in which the core cooling cannot be guaranteed.

Should it be the case, the loss of core coolability engenders a chained sequence of physical phenomena which can end up in core meltdown and the dispersion of contaminants into the environment and the ground. A typical sequence can be as follows: the fuel cladding is oxidized by the steam, which generates hydrogen in the containment; the cladding loses its integrity, and a large part of the fission products is released into the vessel and, through the

circuit and the breach, reaches the containment; the cladding and the fuel lose their geometrical integrity, disaggregate and fall down to a colder region of the core, so that molten "corium" (mixture of core molten materials) is contained inside a solidified crucible; the crucible breaks and the corium falls down into the vessel bottom; if no extra cooling is available after a time, the vessel bottom breaks and the corium falls down and spreads over the basemat of the containment; depending on the chemical, geometrical and thermal conditions, the corium can be either confined and cooled down in the containment, or erodes the basemat and flows down to the ground; the hydrogen in the containment could generate severe damage if its concentration is such that it can cause either detonation or fast deflagration (suitable devices which ignite it as soon as it expands can be added to prevent and mitigate such events); eventually, in case of loss of integrity of the containment, the fission products may be released to the environment, the rate of released radioactivity depending on all the physical-chemical processes that may affect the fission products in the reactor circuits and containment.

All the phenomena involved in a severe accident scenario being very complex and quite coupled, a great difficulty for modeling arises from the lack of precise knowledge of the laws governing them, notably the dynamics of the great number of physical-chemical reactions.

Suitable integral codes have been developed in recent years to perform realistic studies on the accidental scenarios, also - at least partially - accounting for their probabilistic aspects. A typical example of such move is the ASTEC code (Van Dorsselaere et al., March 2009), jointly developed by IRSN and GRS (Gesellschaft für Anlagen- und Reaktorsicherheit mbH), and assessed by 30 organizations in the framework of the SARNET Network of Excellence dedicated to Severe Accidents (Micaelli et al., 2005) and backed by the European Commission in the 6th and 7th Framework Programs. The code is now considered as the European reference for severe accident analysis.

Such integral codes describe all the physical phenomena governing the reactor behavior, in space and time, from the core melting up to the possible release of contaminants to the environment, as well as the behavior of all safety systems and of the operators' procedures (see the scheme of ASTEC code in Figure 5). They must be (relatively) fast running to enable sufficient number of simulations of different scenarios to be performed, accompanied by studies on the uncertainties and on potential cliff-effects. In most codes, the structure is modular enough in order to make easier the validation process, for instance applying only a limited set of modules on experiments devoted to a few physical phenomena (see Fig. 5 for the modular structure of the ASTEC code). As the integral code approaches emphasize the overall plant response, interactions and feedback between separate phenomena occurring at the same time play an important role: e.g. fluid flows, heat transfers, phase changes (melting, freezing, vaporization) and chemical reactions. Another important feature of such codes is that they gather very diverse scientific domains like thermal-hydraulics, chemistry, mechanics of solid structures, neutronics, etc.

Each phenomenon is represented through simplified models, often empirically adjusted on experiments. These codes are globally assessed on integral tests such as those carried out within the PHEBUS FP programs (Clément et al., 2003). Some specific parts of the accident are addressed via the so-called mechanistic codes, which model the local equations more precisely, with a much refined geometrical description. Such codes calculate the behavior of both the core during the degradation process and the corium molten pool in the bottom of

the vessel and in the cavity, as well as the steam and hydrogen distribution in the reactor containment.

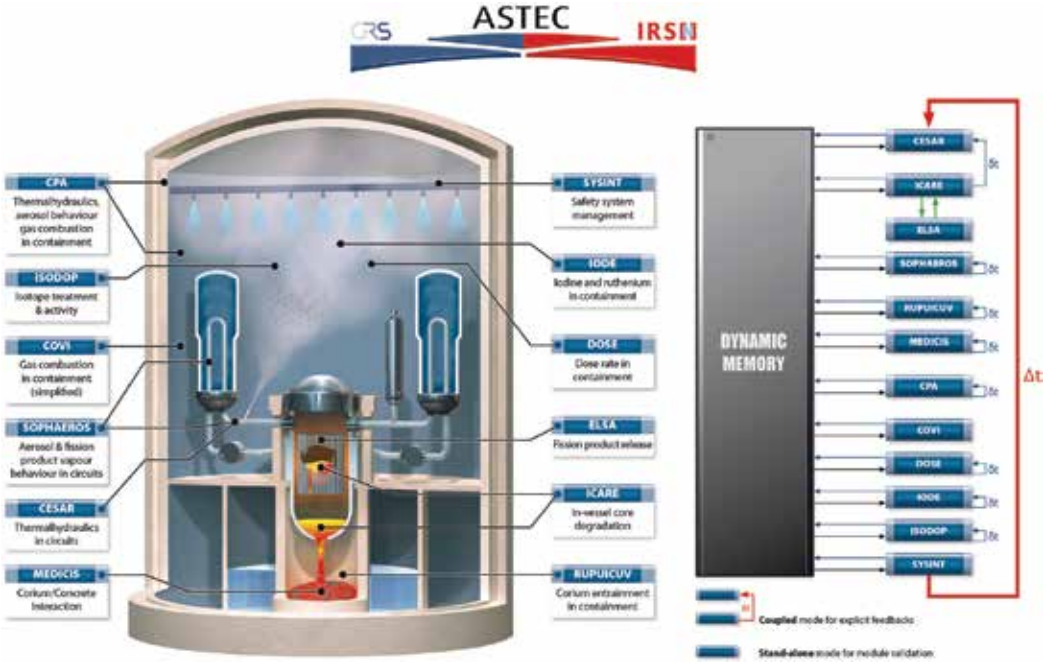


Fig. 5. The ASTEC integral code for simulation of severe accidents

The majority of these codes is still making strong approximations on the geometry of the core and its evolution during the degradation process, and remains very sensitive to the physical-chemical uncertainties, due to the large number of components in interaction and the very high temperature. Such insufficient mastering of the uncertainties, which is at least partially due to the poor knowledge of the behavior and properties of the materials, does not encourage going through further development of very detailed models considering that the outcome of the development efforts should show up quite low.

Nevertheless, selected efforts could be devoted to improving the computation features of the transient parts where the geometry-related effects are widely dominant on the physical uncertainties. In such situations, CFD codes can give interesting results, and in fact they are becoming more and more widely used in problems like the hydrogen repartition and combustion in the containment (TONUS code jointly developed by CEA and IRSN (Bielert et al., 2001)) or the corium pool behavior in the vessel bottom, or the corium spreading and solidification process out of the vessel (CROCO code developed by IRSN (Gastaldo et al., 2006)).

The multi-scale approach has tentatively been adopted at IRSN to investigate the physical phenomena in the regions where the solid particles form a porous bed of debris (Fichot et al., 2006) and where the molten materials build up and accumulate, forming a molten pool (Roux et al. 2006). The progression and growth of the molten pool is a major threat for the vessel wall and, therefore, an important source of concern for the safety experts. One of the most efficient ways to stop its growth is to re-flood it with water but this process involves complex steam and water flows through the porous debris bed. It also forms a solidification

front at the edges of the molten pool where coupled heat transfer and material transport engender major modeling difficulties which challenge the validity of simple models.
The details of the processes (steam and water flow in porous debris and solidification of molten mixtures) should be studied numerically at a small scale; models suitable for implementation in industrial codes could then be derived thanks to a volume averaging method (Fichot et al., 2006; Roux et al., 2006). As previously mentioned in section 3.2, such an approach does not exempt from validating the codes against experiments that involve simultaneously all the phenomena contributing to the process to be modeled. For this reason, in the framework of SARNET, IRSN and partner organizations are building an experimental program that addresses the issue of debris bed quenching by water injection (Van Dorsselaere et al., October 2006).

3.5 Use of CFD codes for other accident studies

Some transients, even if not explicitly included in the set of severe accident initiators, may have important safety consequences and must therefore be studied very carefully. That is the case for the reactivity swing resulting from the injection of clear water into a core at shutdown for reloading. The core is under-critical in these conditions, due to the huge soluble boron poisoning of the water. The injection of clear water generates a RIA–type transient and the core can go back critical quite quickly (and, maybe, even prompt-critical, depending on the clear-water injection amount and location). Immediately, the power of the core begins to increase and it still does until the Doppler feedback is able to shut the reactor down. Then, the cooling-down can start a reactivity-driven oscillation.
Past studies showed that such situations, due to operation and maintenance errors, may be quite likely and significantly contribute to the risk space. Operating procedures were modified to reduce the probability of such events, and probabilistic safety analyses were performed to evaluate their consequences. Nevertheless, they remain a major safety issue and have to be conveniently addressed through computation.
A typical event of this kind is as follows: one of the loops of the reactor provides pure water and the other loops provide water with normal boron concentration level. The main modeling problem is to evaluate the map of boron concentration at the core entry, accounting for the fact that the flow entering into the vessel is highly turbulent, and there are many obstacles opposing the flow, such as tubes and plates in the vessel bottom. A neutron dynamics code can then calculate the core power distribution and evolution with time. Calculations of that kind are already performed with CFD computer codes. To gain full confidence and access to fully realistic results, they need improvements in turbulence models and geometrical modeling, which implies the use of high computing power.
Other studies of operational transient adopt CFD techniques to complement the usual tools and obtain a more precise description of local and complex phenomena such as the flow stratification in pipes and tees, the cold plumes touching hot walls, the impinging jets with temperature differences and the pressurized thermal shocks.

4. Advanced numerical simulation and safety demonstration of GEN IV concepts

Specific needs in terms of development and assessment of advanced computation tools could show up for each GEN IV design, depending on its physical features and operating mode. Nevertheless, several trends can be pointed out as relevant to the safety

demonstration and widely independent from the design. They would claim a major effort of computer code development and assessment, which should impulse new experimental programs.

As mentioned above, particular care is paid in this paper to three out of the six GEN IV concepts:

- The SFR that can benefit from a significant experience in France, Great-Britain, Japan, Russia (and some other countries of the former USSR) and the USA,
- The GFR that presents a very high potential in terms of uranium sparing, incineration, transmutation and heat production; however, even if the concept principles are not new in Europe or in the USA, no GFR has ever been built in the world,
- The HTR/VHTR that can benefit from a first experience in Germany, Great-Britain, China, Japan and the USA.

At the present stage of the investigation of the 3 above-mentioned concepts, five main issues have been retained by IRSN as major ones:

- The consistence and robustness of neutronics design of such systems, the behavior of which is quite different from current PWRs and conventional experimental facilities, due to an increased coupling among neutron and temperature fields, the new design of the core, with heterogeneities, an advanced fuel technology, and a very different operation mode;
- The demonstration of the actual capacity of such systems to passively and safely evacuate the residual power, in any circumstance;
- The features of reactor fuel, with specific emphasis on its transient behavior, mainly as regards either the TRISO particle for HTRs/VHTRs or the advanced carbide and nitrite fuels for fast neutron reactors;
- The features of the source term produced by the migration of activated fission products inside the reactors and likely to be released to the environment in case of accidental situations;
- The inquiry upon either the significant reduction or the risk of a generalized and severe damage of the core, which founds the whole safety approach for these plants.

All these issues are widely addressed in the SRA (Strategic Research Agenda) of the European SNETP Sustainable Nuclear Energy Technology Platform (Bruna et al., 2009) and in several connected presentations and articles, such as (Bruna, 2008). They will not be investigated here. In the following, only the computation-related aspect will be discussed, mainly in the perspective of the improvements expected from either an extended use or the adoption of the CFD methodologies.

All these fields claim for a new effort in R&D. In order to achieve an optimum management of the resources, a priority scale is to be established in agreement with the technological choices and the objective dictated by each country's policies. In the following, we shortly assess each of them before focusing on specific needs for the safety demonstration of the systems which are most likely to be constructed in a relatively near future. It is remembered, for completeness sake, that numerical simulation for the development of specific non destructive examination methods is not addressed in this chapter.

4.1 Reactor physics and core design

GEN IV reactors are very different from each other as regards neutron design, core physics and operating mode. They span a very large spectrum of configurations, including small

and large size cores, fast-neutron and moderated ones, gas, water and liquid metal cooled systems, each one matching more or less completely and comprehensively the general objectives of GEN IV. Sustainability and actinide transmutation are the most affordable goals for systems with fast neutron flux, such as SFRs and GFRs. On the contrary, graphite-moderated gas-cooled thermal-flux reactors, such HTRs and VHTRs, are most likely to be inherently safe and to allow a diversified energy production (electricity, but also industrial steam and hydrogen).

In addition to the overall design, the core size and the operating modes, the fuel, the materials for internals and vessel, the coolant features generate specific problems which must be assessed in computations. Moreover, a strong coupling among neutron and temperature fields can show up in large-size systems. Simulation challenges can be sharpened by the coupling with conventional energy production systems, which can propagate instability and perturbation to the reactors, through the intermediate heat exchanger.

Accordingly, the requirements in terms of simulation for core physics and operation studies would be quite different. A sometimes massive heterogeneity in space and energy and the mutual interactions between the neutron and temperature fields claim for new and enlarged 3D capabilities, and an increased coupling for design and normal operation calculations.

Integrated systems permitting a full description of coupled neutronics, thermal and mechanical transients, such as the SIMMER III/IV code (Tobita et al., 2006), should be very useful for safety studies of strongly coupled, fast-kinetics systems, such as SFR and GFR systems. On the other hand, for HTR and VHTR systems, due to the strong dependence of the core equilibrium on the temperatures, focus should be put on bulk codes enabling a full coupling among the core and the reflector temperature and neutron fields.

Moreover, specific needs exist for SFRs, which mainly concern the risk of a generalized and severe damage of the core, due to either reactivity-driven transients, such as the coolant void (mainly the sodium), or mechanically-initiated transients, such as the blockage of a coolant in a subassembly.

Last but not least, a major safety concern for PBMR (Peddle Bed Modular Reactor) type reactors (particular type of HTR) is the confidence in the evaluation of power peak within an heterogeneous core, where neither the local composition nor the lattice is precisely known during reactor operation, due to the stochastic distribution of the pebbles and the wide burn-up spread among them. Specific developments are needed, which involve a massive use of probabilistic techniques and a careful appreciation of uncertainties. All these items claim for a strong R&D effort devoted both to code development, qualification and validation and to measurement campaigns in ad hoc mock-up experiments.

So as to manage resources as best as possible, a priority scale must be established in agreement with the political and technological choices: emphasis should be put on each item according to its relevance to the safety demonstration of the forerunning concepts likely to be industrialized in a near future.

4.2 Residual power evacuation

For GEN IV concepts as for many other existing ones, the verification of the sufficient cooling of the core in various accidental situations is one of the most important tasks of the safety analysis. Such verification should be supported by the numerical simulation of two processes:

- Fuel cooling by liquid (e.g. sodium) or gas (e.g. helium) natural convection and heat radiation;
- Heat evacuation by water safety circuits.

The difficulty will of course depend on the design (complexity, safety margins, etc.). However, we could reasonably consider that already existing tools like thermal-hydraulics system codes (already developed for light water reactors) and CFD codes for more local evaluation (with radiation models) should be sufficient. Adequate design-oriented experiments will surely have to be performed in order to assess the codes validity for some specificity of the circuits, but this will remain in a strict continuation of current actions aiming at improving capabilities of thermal-hydraulics codes and extending CFD use in reactor safety analysis.

4.3 Fuel integrity

As already mentioned, the integrity of reactor fuel will be an important issue for GEN IV concepts. The challenge will be comparable to that encountered with current generation ones. With a view to enforcing the demonstration of the robustness of the fuel and its resistance to the operation and accidental transients, improvements and adjustments will have to be made in computation tools and devoted experimental programs developed for physical assessment and qualification needs. According to the fuel features and design, it is straightforward that such updating and experiments should be reactor concept-oriented.

The larger effort is foreseeable for HTR/VHTR concepts which, despite their ancient design, have accumulated a quite limited operating experience and, far more, for GFRs, the fuel design of which is new (and is an essential source for performance improvement in terms of both operation and safety, through the achievement of ISO-generation conditions) and does not benefit from any operation feedback.

However, the simulation strategy should be the same as for the current reactor generations: simplified models shall be derived for industrial and well assessed simulation tools and the derivation of these models shall be backed up by a multi-scale approach. It could be recommended to put in place such a strategy as soon as possible in order to more efficiently define the experiments against which the elementary and global assessment of models will be performed.

4.4 Fission products release

All the phenomena involved in the transfer of fission products from the fuel elements to the containment and from the containment to the environment are very complex. As for the current generation of reactors, difficulties come from the great number of involved physical-chemical reactions that make a detailed mechanistic approach almost impossible.

Since the risk of a severe accident and of significant fission products release should be lowered for GEN IV concepts, it does not appear as a necessity to significantly increase the precision we have today when predicting the potential consequences of a severe accident.

Thus, for this topic, it is not judged necessary, from the safety point of view, to have any breakthrough in terms of modeling, apart from the necessity to develop specific models of fission products release for some GEN IV fuels (TRISO for HTRs/VHTRs, carbide for SFRs, specific fuel for GFRs). Simplified models should be sufficient although they will have to be assessed against an appropriate experimental data base including separate effect tests and integral effect tests to make sure that no major important phenomenon has been forgotten.

However, as for the simulation of severe accidents in the current reactor generation, it could be recommended to follow up the current strategy and back up the simplified models by detailed models when it is possible.

4.5 Reduction of the major risk of generalized and severe core damage

As regards the problem of the exclusion of transients likely to result in core melting, it is quite obvious that concepts such as the HTR/VHTR are much more inherently protected against high fuel damaging than others, such as the SFR and the GFR, due to a far slower kinetics, a wider thermal inertia (due to the huge amount of graphite), a capacity to passively evacuate residual heat in almost any circumstance, and a high thermal robustness of the fuel particles.

However, even if the designers' target is to make a whole core melting or high damaging highly hypothetical, a wise strategy would be, in particular for SFRs and GFRs, to investigate

- the mechanisms that could prevent a core local meltdown from degenerating into a whole core meltdown,
- the consequences of a whole core meltdown on containment integrity (including the release of radioactive elements into the environment).

Codes based on simplified models have been developed and used for the previous generations of reactors (LWRs and SFRs). Appropriate experimental programs have been initiated in the 80s to assess these models. The question of the adequacy of these codes and of their assessment for GEN IV concepts can be considered as an open one. It is likely that codes already developed for previous generations of SFRs will be applicable to GEN IV SFRs, provided some complementary developments and assessment are done (the demonstration on core re-criticality risk was not easy and will not be easier for GEN IV, the demonstration of corium retention, etc.).

The adaptation of LWR codes to HTR/VHTR concepts seems possible although, as the core materials are significantly different, all the elementary models will have to be revised and reassessed against a new and appropriate experimental data base.

Phenomena involved in a severe accident are and will remain very complex due to the tight coupling among several phenomena that intervene as driving ones at different instants of the transients: multiphase flows, heat and mass transfers, thermo-chemistry, mechanic resistance of metallic structures, material melting and freezing, core physics and neutron kinetics, etc.

This complexity makes it nearly impossible to envisage in the coming twenty years any revolution in the numerical simulation of these accidents and the conclusions of section 3.4 for LWRs should be considered valid for GEN IV concept severe accidents: the use of advanced numerical simulation could be introduced by CFD or DNS computation in realistic geometries, for calculation of basic averaged values or limited parts of the accident, in support to "integral" codes based on simplified models such as those adopted for the current generations of severe accident codes.

5. Conclusion

Almost all the codes developed during the last twenty-year period for the analysis of the safety problems of nuclear reactors in operation adopt simplified geometry descriptions and

quite simple physical models, stressing the major physical phenomena in some detail only, and either addressing in a quite approximate way or even neglecting the minor ones, as it is the case for the LWR LOCA codes.

That is undoubtedly a drawback to be overcome from the performance point of view since it implies the adoption of operating and safety margins at any stage of the reactor design and operation. Nevertheless, no major changes are expected in the near future as far as the safety analysis of current reactors is concerned, mainly because the computation systems currently in use benefits from a large validation against a set of diversified and extended experimental results. Moreover, the industrial safety applications need to rely on methods agreed by the safety expert organizations. Quite a long time is therefore generally needed before the advanced methods developed by researchers can be adopted in practice to address actual safety cases.

Advanced simulation is undoubtedly able to provide extended capability to calculate local parameters and, accordingly, it allows deeper insights in many problems, contributes to a better understanding of the physics, and thus leads to more reliable designs, reduced costs and/or more precisely quantified safety margins. For system analysis, advanced simulation has thus a complementary role to play in nuclear safety applications in combination with system codes, particularly in those areas where multi-dimensional aspects are relevant. Moreover, combined applications, supported by proper experiments may guarantee a more precise evaluation of safety margins.

Single-phase CFD applications are already reasonably mature although some models (e.g. turbulence and combustion) need improvements. Two-phase and multi-phase CFD modeling still require considerable research efforts even though some aspects may be already reasonably well addressed through the advanced models. In addition, a lot of work in terms of experimentation, model development and assessment has still to be done before practical applications in nuclear safety studies can be made. Thus, as far as the current reactor safety analysis is concerned, the adoption of CFD techniques should mostly be limited to achieving a more detailed understanding of the physical phenomena and supporting the methodology currently in use rather than to supporting the development of fully new computation systems.

Multi-scale techniques are more and more used to consolidate the physical bases of simplified models. The use of these techniques allows progressing more rapidly in the understanding of physical processes and contributes to optimizing experimental programs. However, these techniques are applicable for a limited number of phenomena; they provide models that shall be globally assessed against integral experiments.

On the other hand, as for incoming GEN IV concepts, even if it is assumed that the development pace of computing power keeps constant, due to the complexity of the phenomena and the wideness of the investigation fields, a significant breakthrough in the development of computational tools dedicated to the safety demonstration seems quite unlikely in the short term (roughly within 10 to 15 years).

Moreover, the preliminary studies of some of these concepts now underway allow believing that there is no specific need for profound modifications to the current code development and assessment strategies. The key element will remain the adequate validation of the computation chains against appropriate analytical and integral tests, which means *in fine* uncertainty and design margins.

Accordingly, it seems likely that large improvements of computation tools, including an extensive adoption of CFD methodologies, are scheduled for the intermediate future (within the next 25 years).

Nevertheless, it is likely that the use of advanced modeling will be extended and reinforced for GEN IV fields of endeavor, alongside with the expansion of the application for current reactors.

The most challenging issues in the methodology to GEN IV computation should be:

- The extended adoption of CFD techniques for single-phase application, addressing core cooling in particular, with the support of some specific experimentation assessing models and calculation methodologies;
- The development of multi-physics computational tools with a tight coupling among core physics, fuel thermal-hydraulics and thermo-mechanics, as well as systems description;
- The increase in the predictability of fuel codes for fuel integrity issues. To comply with the expected continuous process of fuel improvement, this should be backed by the development of a multi-scale strategy (already initiated for LWR fuel) and supported by a suitable experimental activity as well;
- The achievement, in the severe accident and source term evaluation issues, of a modeling level close to that achieved for LWRs, for which the advanced modeling is only seen as a support for a better understanding of some physical aspects of involved phenomena.

As a general conclusion, in the present state of knowledge, no major breakthroughs seem necessary in terms of modeling for reactors in operation, at least whether if it is postulated that no significant changes are adopted in their design features and operation. Simplified models should still be satisfactory enough, provided that they are validated on appropriate and representative experimental data, including results from both analytical and integral tests.

As far as the future GEN IV concepts are concerned, it must be emphasized that the current wide effort for updating models should provide opportunity for "boosting" advanced numerical simulation that is undoubtedly a source for better understanding of the system physics and consequently improving the concept design and future operation. Nevertheless, the safety analysis being strictly dependent on reactor design, a further investigation on this relevant topic is to be carried out once the main design and operation options for those systems is definitely known.

6. References

Bielert U., Breitung W., Kotchourko A., Royl, P. Scholtyssek W. , Veser A. , Beccantini, A. Dabbene F. , Paillere H., Studer E., Huld T., Wilkening H., Edlinger B., Poruba C. and Mohaved M. (2001) Multi-dimensional simulation of hydrogen distribution and turbulent combustion in severe accidents, *Nuclear Engineering and Design,* Volume 208, Issues 1-3, Pages 165-172, November 2001

Bruna G.B., Fouquet F., Dubois F., Le Pallec J.-C., Richebois E., Hourcade E., Poinot-Salanon C. and Royer E. (2007) HEMERA: a 3D Coupled Core-plant System for Accidental Reactor Transient Simulation, *Proceedings of the ICAPP 2007, International Meeting,* Nice, Acropolis, France, May 13-18, 2007

Bruna G.B. et al. (2008) SNETP – Sustainable Nuclear Energy Technology Platform - Strategic Research Agenda - SRA 2009, www.SNETP.eu, issued, May 2009

Bruna G.B. (2009) The Vision of the European Sustainable Nuclear Energy Technology Platform Strategic Research Agenda on the Safety R&D for GEN-IV Reactors, invited panel paper at the Bulgarian Nuclear Society 2008 Conference "Nuclear Power for the People. Nuclear Installation Safety and Environment", Sofia, November 2008, Published in the Science and Technology Journal, Vol. 13, N° 2, September 2009, BgNS Transactions, ISSN 1310-8727

Clergeau. M, Normand B., Sargeni A. (2010) HEMERA: 3D computational tool for analysis of accidental transients, Proceedings of the *PHYSOR 2010 Meeting – Advances in Reactor Physics to Power the Nuclear Renaissance,* Pittsburgh, Pennsylvania, USA, May 9-14, 2010, on CD-ROM, American Nuclear Society, LaGrange Park, IL 2010

Clément B., Hanniet-Girault N., Repetto G., Jacquemain D., Jones A.V., Kissane M.P. and von der Hardt P. (2003) LWR severe accident simulation: synthesis of the results and interpretation of the first Phebus FP experiment FPT0, *Nuclear Engineering and Design*, Volume 226, Issue 1, Pages 5-82, November 2003

Fichot F., Duval F., Trégourès N., Béchaud C. and Quintard M. (2006) The impact of thermal non-equilibrium and large-scale 2D/3D effects on debris bed reflooding and coolability, *Nuclear Engineering and Design*, Volume 236, Issues 19-21, Pages 2144-2163, October 2006

Gastaldo L., Babik F., Herbin., R. and Latché J.-C. (2006) An unconditionally stable pressure correction scheme for barotropic compressible Navier-Stokes equations, *Proc. ECCOMAS CFD 2006, s, TU Delft, Netherland*

Livolant M., Durin M. and. Micaelli J.-C. (2003) Super Computing and Nuclear Safety, *Proceedings. of the International Conference on Supercomputing in Nuclear Applications, SNA 2003,* Paris, France September 22-24, 2003

Micaelli J.-C., Haste T., Van Dorsselaere J.-P., Bonnet, J.M., Meyer L., Beraha D., Annunziato A., Chaumont B., Adroguer B., Sehgal R. and Trambauer K. (2005) SARNET: a European cooperative effort on LWR severe accident research, *Proceedings of ENC 2005,* Versailles, France, December 11-14, 2005

OECD IAEA (2002) Use of Computational Fluid Dynamics Codes for Safety Analysis of Reactor Systems, including Containment, *Minutes of the Technical Meeting,* Pisa, Italy, November 11-14, 2002

OECD/NEA (2003) Halden Reactor Project, *http://www.nea.fr/html/jointproj/halden.html.*

Papin J., Cazalis B., Frizonnet J.M., Desquines J., Lemoine F., Georgenthum V., Lamare F. and Petit M. (2007) Summary and Interpretation of the CABRI REP-Na Program, *Nuclear Technology,* Volume 157, N°3, pages 230-250, March 2009

Perales F., Monerie Y., Chrysochoos A., (2006) Non-smooth fracture dynamics of functionally graded materials, *Journal of Physics IV, Volume 134, pages 367-372,* 2006

Perales F., Bourgeois S., Chrysochoos A., Monerie Y. (2008) Two-field multi-body method for periodic homogenization in fracture mechanics of non linear heterogeneous materials, *Engineering Fracture Mechanics,* Volume 75, pages 3378-3398, 2008

Roux P., Goyeau, B. Gobin D., Fichot F. and Quintard M. (2006) Chemical non-equilibrium modeling of columnar solidification , *International Journal of Heat and Mass Transfer*, Volume 49, Issues 23-24, Pages 4496-4510, November 2006

Schmitz F. and Papin J., (1999) High burn-up effects on fuel behavior under accident conditions: the tests CABRI REP-Na, *Journal of Nuclear Materials*, Volume 270, Issues 1-2, pages 55-64, April, 1, 1999

Suzuki, M. Saitou H. and Fuketa T. (2006) Analysis on split failure of cladding of high burn-up BWR rods in reactivity-initiated accident conditions by RANNS code, *Nuclear Engineering and Design*, Volume 236, Issue 2, Pages 128-139, January 2006

Tobita Y., Kondo Sa., Yamano., H., Morita. K., Maschek W., Coste P. and Cadiou T. (2006) The Development of SIMMER-III, An Advanced Computer Program for LMFR Safety Analysis, and Its Application to Sodium Experiments, *Nuclear Technology,* Volume 153, Number 3, Pages 245-255, March 2006

Van Dorsselaere J.P, Fichot F. and Seiler J.-M. Views on R&D needs about in-vessel reflooding issues, with a focus on debris coolability, *Nuclear Engineering and Design*, Volume 236, Issues 19-21, Pages 1976-1990, October 2006

Van Dorsselaere J.P., Seropian C., Chatelard P., Jacq F., Fleurot J., Giordano P., Reinke N., Schwinges B., Allelein H.J., Luther W. The ASTEC integral code for severe accident simulation, *Nuclear Technology*, Volume 165, Pages 293-307, March 2009

Edited by Lutz Angermann

This book will interest researchers, scientists, engineers and graduate students in many disciplines, who make use of mathematical modeling and computer simulation. Although it represents only a small sample of the research activity on numerical simulations, the book will certainly serve as a valuable tool for researchers interested in getting involved in this multidisciplinary field. It will be useful to encourage further experimental and theoretical researches in the above mentioned areas of numerical simulation.

ISBN 978-953-51-5555-3

Photo by gmac84 / iStock

IntechOpen